DOT/FAA/AM-95/24

Office of Aviation Medicine
Washington, D.C. 20591

# The Effect of Alcohol and Fatigue on an FAA Readiness-To-Perform Test

NTI, Inc.
Dayton, Ohio 45432

Civil Aeromedical Institute
Federal Aviation Administration
Oklahoma City, Oklahoma 73125



August 1995

Final Report



U.S. Department of Transportation
Federal Aviation Administration

# NOTICE

**Technical Report Documentation Page**

| 1. Report No. | 2. Government Accession No. | 3. Recipient's Catalog No. |
|---|---|---|
| DOT/FAA/AM-95/24 | | |

| 4. Title and Subtitle | 5. Report Date |
|---|---|
| The Effect of Alcohol and Fatigue on an FAA Readiness-To-Perform Test | August 1995 |
| | 6. Performing Organization Code |
| 7. Author(s) | 8. Performing Organization Report No. |
| NTI, Inc. | |
| 9. Performing Organization Name and Address | 10. Work Unit No. (TRAIS) |
| NTI, Inc.<br>4130 Linden Ave., Suite 235<br>Dayton, OH 45432 | 11. Contract or Grant No.<br>DTFA 01-93-C-00004 |
| 12. Sponsoring Agency name and Address | 13. Type of Report and Period Covered |
| Office of Aviation Medicine<br>Federal Aviation Administration<br>800 Independence Ave., S.W.<br>Washington, DC 20591 | Final Report |
| | 14. Sponsoring Agency Code |

15. Supplemental Notes

This work was performed under contract from the Human Resources Research Division of the FAA's Civil Aeromedical Institute (CAMI) in Oklahoma City, Oklahoma, as part of task AM-8-93-HRR126. Thomas E. Nesthus, Ph.D., served as CAMI's research technical representative.

16. Abstract

Readiness to perform (RTP) testing is considered by some to be a broad-based alternative or supplement to biochemical testing for drugs and alcohol. Since it is also thought to detect impairment due to other sources (e.g., fatigue, illness, depression), the Federal Aviation Administration (FAA) is interested in exploring its scientific validity and practical utility. This study defined the statistical sensitivity and individual diagnosticity of an RTP test utilizing the NovaScan™ paradigm. METHODS: 77 male subjects within 3 age groupings (25-34, 40-48, and 54-62) were administered alcohol sufficient to raise their breath alcohol content (BrAC) to .08% BrAC. FAA-NovaScan testing occurred once each hour as their BrAC levels rose to .08% and diminished back to baseline levels. The double-blind design involved having alcohol drinks and "sham" alcohol drinks administered in a counter-balanced order on 2 separate days. RESULTS: An estimate of the "reliability" of the test once it reached plateau levels indicated that most reaction time variables had a reliability between .76 and .94, with some percent correct measures showing too little variability to calculate meaningful reliabilities. Multivariate and univariate analyses of variance were conducted to determine whether the test was sensitive to various levels of BrAC. Ingestion of alcohol produced statistically significant effects on RTP test performance. Reaction time measures on all 3 tasks in the FAA-RTP test showed statistically significant decrements during the alcohol ingestion phase of the alcohol day that were monotonically related to BrAC level while BrAC was increasing. When BrAC was decreasing, alcohol-induced decrements were generally more severe, and were not monotonically related to the BrAC levels in all cases. A task requiring repetitive attention appeared most sensitive to alcohol concentration, followed by a task requiring mental rotation and memory. A visual search and memory task, although not as effective in detecting alcohol levels, showed some significant effects, apparently contributing to the efficiency of the entire test. Candidate scoring algorithms were developed to determine whether the test could have detected individuals at each BrAC level. When cut-off points of 2.0 standard deviations were used on several test variables, the procedure would have detected 97% of the subjects at .08% BrAC, 88% at .06% BrAC, and 76% at .04% BrAC. With this criterion, 30% of the subjects would also have "failed" the test, even with no alcohol in their system. Inspection of results on the placebo day revealed that when the test was administered twice, as it would in actual implementation, this false positive rate was reduced to 24%. CONCLUSIONS: The FAA-RTP test is sensitive in detecting performance decrements due to the generally accepted levels of legal alcohol intoxication. As such, it shows promise as a non-invasive screening procedure.

| 17. Key Words | | 18. Distribution Statement |
|---|---|---|
| Performance-Based Testing | Attention Switching | Document is available to the public through the National Technical Information Service Springfield, Virginia 22161 |
| Readiness-To-Perform | Fitness-For-Duty | |
| Validation Test | Alcohol | Fatigue |

| 19. Security Classif. (of this report) | 20. Security Classif. (of this page) | 21. No. of Pages | 22. Price |
|---|---|---|---|
| Unclassified | Unclassified | 68 | |

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

## ACKNOWLEDGMENTS

This report documents work performed by NTI, Inc. for the Federal Aviation Administration under contract DTFA-01-93-C-00004. The principal investigator for this effort was Dr. R. O'Donnell, and critical technical assistance was provided by Dr. S. Moise, Ms. D. Warner, Ms. R. Cardenas, Ms. D. Krapff, and Mr. R. O'Donnell, all of the NTI staff.

The authors wish to thank Dr. J. Brecht-Clark for her insight and technical skill in conceptualizing and implementing this project. In addition, Dr. T. Nesthus provided invaluable technical direction and assistance in carrying the project to completion. Mr. H. Harris and Mr. M. Touchstone provided outstanding support to the project in administering the alcohol to subjects. Their expertise is documented in these pages, but their dedication and professionalism can only be appreciated by those who were involved. Dr. K. Gilliland and Dr. R. Schlegel provided early reviews of the work which significantly enhanced the scope of the analyses. In addition, the authors are indebted to Dr. D. Taylor, Dr. D. Schroeder, Dr. D. Broach, and Dr. R. Blanchard of the FAA for their careful and insightful comments on earlier drafts of this report.

## TABLE OF CONTENTS

## TABLE OF CONTENTS (CONTINUED)



## LIST OF TABLES



## LIST OF FIGURES

## EXECUTIVE SUMMARY

Readiness to perform (RTP) testing is considered by some to be a broad-based alternative or supplement to biochemical testing for drugs and alcohol. Since it is also thought to detect impairment due to other sources (e.g., fatigue, illness, depression), the Federal Aviation Administration is interested in exploring its scientific validity and practical utility. The present study represents the first effort to define the statistical sensitivity and individual diagnosticity of an RTP test developed specifically for the FAA utilizing the NovaScan™ paradigm.

A total of 77 male subjects within specified age groupings were administered alcohol sufficient to raise their breath alcohol content (BrAC) by .02% per hour, up to a limit of .08% BrAC. They were tested with a new FAA-RTP test once each hour as their BrAC levels rose to .08% and diminished back to baseline levels. Over 9 hours of testing, subjects also performed a visually demanding, but entertaining video game, in order to simulate a work day. The double-blind design involved having alcohol drinks and "sham" alcohol drinks administered in a counter-balanced order on 2 separate days.

The median number of training trials before subjects reached plateau performance on most test variables was 16.5 sessions, with the 90th percentile point being 23.4 sessions. Based on this, and on estimates of the time necessary to take the test, it was concluded that the median subject would require approximately 2.75 hours of total training time to reach plateau, and that 90% of this type of subject will reach plateau with a total training time of 3.9 hours. An estimate of the "reliability" of the test once it reached plateau levels indicated that most reaction time variables had a reliability between .76 and .94, with some percent correct measures showing too little variability to calculate meaningful reliabilities.

Analyses of results on the "fatigue" (placebo) day revealed a subjectively increasing level of fatigue over the course of a day. However, this was not reflected in the average curves of the subjects for RTP variables. Thus, although subjects were experiencing more subjective fatigue as the day wore on, their level of performance did not deteriorate to statistically significant levels.

RTP performance scores were subjected to multivariate and univariate analyses of variance to determine whether the test was sensitive to various levels of BrAC. Ingestion of alcohol produced statistically significant effects on RTP test performance. Reaction time measures on all 3 tasks in the FAA-RTP test showed statistically significant decrements during the alcohol ingestion phase of the alcohol day that were monotonically related to BrAC level while BrAC was increasing. When BrAC was decreasing, alcohol-induced decrements were generally more severe, and were not monotonically related to the BrAC levels in all cases. A task requiring repetitive attention appeared most sensitive to alcohol concentration, followed by a task requiring mental rotation and memory. A visual search and memory task, although not as effective in detecting alcohol levels, did show some significant effects, and did appear to contribute to the efficiency of the entire test.

In addition to the statistical analysis of sensitivity, individual analyses of subjects' performance related to alcohol level were carried out. Candidate scoring algorithms were developed to determine, on an individual basis, whether the test could have detected individuals at each BrAC level (i.e., whether subjects would have "failed" the test at each BrAC level). When cut-off points of 2.0 standard deviations were used on several test variables, the procedure would have detected 97% of the subjects at .08% BrAC, 88% at .06% BrAC, and 76% at .04% BrAC. With this criterion, 30% of the subjects would also have "failed" the test, even with no alcohol in their system (false positives with respect to alcohol). Inspection of results on the placebo day revealed that when the test was administered twice, as it is intended to be used in actual implementations, this false positive rate was reduced to 24%. A more stringent cut-off criterion (-1.5 standard deviations) resulted in detection of 100% of subjects

at .04% and .08% BrAC, and 90% at .06% BrAC. As expected, this increased detection comes at a cost of a higher "false positive" rate of 44% with 1 test and 41% with 2 tests.

Finally, subjective estimates of the acceptability of the test were obtained by post-test questionnaires. A very high percentage of the subjects (89%) believed that the customized FAA NovaScan test would detect decrements due to alcohol, fatigue, drugs, and over-the-counter medications. Concurrently, a significant percentage of subjects misjudged whether they had been given alcohol or not, with 34% indicating that at no point during the day (even when they were at .08% BAC) would they have chosen not to drive their car.

The above results indicate that the FAA-RTP test developed here is sensitive in detecting performance decrements due to the generally accepted levels of legal alcohol intoxication. As such, it shows promise as a non-invasive screening procedure that could be used as "reasonable cause" for further testing. In view of these results, it is recommended that a criterion-based study be carried out to cross-validate the sensitivity of this type of RTP test in detecting an alcohol stressor, and which also would establish the relationship between performance on the FAA-RTP test and performance in a real-world environment of interest to the FAA. In addition, further study of the logistics, cost, administrative, and legal issues associated with use of this test, and RTP testing in general, appears warranted.

# The Effect of Alcohol and Fatigue on an FAA Readiness-To-Perform Test

## SECTION 1
## INTRODUCTION

The FAA is interested in evaluating the utility and sensitivity of Readiness to Perform (RTP) tests for possible implementation into work force safety-sensitive positions. RTP has been defined as "...that state in which a person is prepared and capable of performing a job for which the person is willingly disposed and is free of any transient risk factors, such as drugs, alcohol, fatigue, or illness, that might influence job performance" (Gilliland and Schlegel, 1993). RTP tests are those designed to evaluate that state, especially on a short-term basis, to detect individuals who may not be ready to safely perform their job. If found to be valid, reliable, and practical, such tests could significantly enhance the already outstanding margin of safety in aviation. This is so because RTP tests, while potentially as sensitive as drug and alcohol testing for detecting those sources of decrement, may also detect sources of performance impairment from a large variety of other causes.

After a broad market survey of available RTP tests, the FAA selected one particular type of RTP test for further study. This was the NovaScan™ test procedure. Generically, NovaScan is a test framework in which the individual is required to process 3 or more tasks in a near-simultaneous way (O'Donnell, 1992). It is presented on a personal computer, and typically takes less than 10 minutes to administer, with each individual tested against his or her personal baseline. The actual tests used in the general NovaScan framework can be tailored to probe skills critical to specific jobs. Early validation studies of various NovaScan implementations indicated that it was sensitive to certain drugs, and to levels of blood alcohol in the .04% to .05% range (O'Donnell, 1993a). Performance on the test correlated highly with job performance in flying, driving, and control-room operation (O'Donnell, 1993b).

A plan was developed to produce an FAA-specific version of the NovaScan test, and to subject it to a series of validation studies to determine its utility in the FAA environment. The present study constitutes the first of these efforts.

A fundamental question requiring resolution deals with the basic sensitivity of the NovaScan procedure to known stressors. Obviously, one of the best-defined of such stressors is alcohol. Although there is no universally agreed upon relationship between low levels of blood alcohol and performance, government agencies have specified acceptable levels of blood alcohol concentration (BAC) for various activities. In most states, a BAC level below .10% is legal for operating a private automobile. In some states, this level is set at .08%. A level below .04% is considered acceptable for operating a commercial vehicle, and for many other safety-sensitive jobs. These levels, since they are well defined and verifiable, constitute a logical starting point for determining the sensitivity of the FAA NovaScan test. For this reason, this first study was aimed at defining the basic statistical sensitivity of NovaScan to alcohol, and its ability to detect individuals who were known to have various blood alcohol concentrations in their system.

## SECTION 2
## METHODS AND PROCEDURES

### Overview

The basic design of this study was a 2-factor, double blind procedure. The first major factor was the presence or absence of various levels of blood alcohol concentration (BAC). BAC was estimated by breath alcohol concentration (BrAC), which is considered to be a reasonable approximation of BAC under proper conditions. The second factor consisted of time-on-task, referred to below as "fatigue." This was induced by having the subjects perform a visually demanding set of tests and tasks over a 9-hour period. The major dependent variable was performance of the subject on a version of the NovaScan test procedure, which was specifically designed for the FAA. This test will be referred to as the FAA Readiness to Perform (FAA-RTP) test in this report.

### Subject Recruiting Procedures

A general recruitment was carried out in the local Oklahoma City, OK area to obtain subjects for this study. A newspaper ad was placed in the *Daily Oklahoman*, presenting the basic nature of the study and initial entrance criteria. Over 200 responses was screened by phone to determine whether they were generally in good health and had no history of alcohol abuse. Based on this screening, a total of 62 respondents were tentatively scheduled for participation in the study. In addition, 25 individuals were recruited from a list of subjects who had previously participated in an alcohol study at the FAA Civil Aeromedical Institute. Finally, 27 subjects were recruited through individual contacts with civic clubs, benevolent organizations, and churches throughout the Oklahoma City area.

In summary, 114 subjects were scheduled for participation. Of these, 18 failed to report for their scheduled training, and 3 withdrew voluntarily for personal reasons after 1 or more training sessions. Further, 8 subjects were eliminated from actual study for medical reasons or because of a positive history of alcohol abuse. This left 85 subjects who were given at least 1 test session. However, 3 of these subjects received only 1 test day (because 2 failed to report for the second day, and 1 reported for the test day with a BrAC of .028, and was not tested on that day). One subject, with a previous history of alcoholism, was tested on 2 days, but was not given any alcohol on either day. Finally, 4 subjects became physically ill during the test (including 1 on the placebo day), and could not complete all test requirements. This left a total of 77 subjects who completed training and both days of testing. The subjects who did not complete testing appeared similar in age, intelligence, and occupation to those completing testing. The 77 subjects actually used in the experiment appear as a reasonably representative sample of the general male population of the southwest portion of the United States.

### Description of the Test Instruments

Six test instruments were used in this experiment. One of these, the Mah Jongg test, was neither scored nor analyzed. This instrument was used as a "filler" to ensure a high level of visual work demand among subjects on test days. Four of the other instruments were self-report scales of 1 form or another. Each provided background data for the experiment. These will be described briefly below. The major dependent variable for the study was performance on the FAA-RTP test, which will be described in considerable detail below.

### The Mah Jongg Game

This is a popularized computer version of the traditional Mah Jongg game. It was obtained from Shareware (developed by Nels Anderson), and was used in unmodified form. The object of the game is to remove as many of the tiles from the playing board as possible, within the rules of the game. The player is instructed to remove tiles in sets of matching tiles, but a tile is available for removal only if either its left or right edge is unblocked by another tile. There are 5 layers of tiles stacked 1 above the other in pyramid fashion, so the player has several edge tiles from which to choose. Subjects in this

experiment were told their scores would be calculated on the basis of 2 criteria: (1) the number of different tile sets he played over the course of the two testing days; and (2) the number of tiles removed from each game board. A complete set of instructions given to the subject for this game is presented in NTI, Inc. (1993).

**Medical Questionnaire**

To assure that subjects were not suffering from any medical condition that could be exacerbated by any of the experimental procedures, a physician developed a basic questionnaire to isolate such conditions (ibid.). During the Human Use protocol review, several additions were made to the basic questionnaire by the medical board members. These were added as questions 10 through 12 of the questionnaire. Any "yes" answer resulted in the case being reviewed by the FAA physician on-call. This procedure resulted in cases being referred to the physician. Of these referrals, 8 subjects were eliminated from the study because of various self-reported medical conditions, including irregular heartbeat, knee injury, long periods of abstinence from alcohol (punctuated by periods of excessive drinking), liver condition, prescriptions requiring medicines that should not be taken in conjunction with alcohol, and alcoholism.

**Fatigue Questionnaire**

This instrument was used to obtain some indication of how the subjects' perceived fatigue level changed during the test days. Subjects were requested to fill out the fatigue form approximately once per hour over the course of each test day. The instrument used was the School of Aerospace Medicine Fatigue Questionnaire. This questionnaire has been used for many years by the U.S. Air Force as a quick subjective estimate of self-perceived fatigue. Essentially, the fatigue questionnaire is a series of 10 statements that the individual rates relative to his current perceived level of fatigue. Such statements as, "extremely peppy" or "ready to drop," are rated as to whether the individual feels the same as the statement, better than the statement, or worse than the statement. Each of these ratings is then given a numerical value, and the sum of all of the ratings is used to estimate the subject's current state of self-perceived fatigue.

**Alcohol Use Questionnaire**

To provide an estimate of the subject's prior alcohol use, a subjective report questionnaire was adapted from a longer questionnaire used by the University of Oklahoma Alcohol Research Unit. Portions of the extensive screening questionnaire used by that unit were extracted (with permission) and filled out by the subjects. Essentially, the questionnaire requests information on the subject's early use of alcohol, his typical current drinking pattern, and the type and potency of alcohol currently used. For the present purposes, interest was in determining whether each individual subject was a heavy drinker, moderate drinker, light drinker, infrequent drinker, or abstainer, according to the criteria established in the "Quantity-Frequency-Variability Index" (Cahalan, Cisin, and Crossley, 1967). In addition, this instrument was used to determine whether subjects may have a problem with alcohol use, since this was an exclusionary criterion for the study.

**The Shipley Institute of Living Scale**

To provide a crude estimate of verbal intelligence level, we had the subjects complete the verbal subscale of the Shipley-Institute of Living Scale for Measuring Intellectual Impairment (Shipley, 1940). In its full form, this scale is reported to correlate 0.85 with the Wechsler Adult Intelligence Scale (WAIS-R), which has a mean of 100 and a standard deviation of 15 (Gregory, 1987; Zachary, Crumpton, and Spiegel, 1985). Even in its full-scale form, the test is only considered to be a rough approximation of true IQ, being designed primarily to eliminate sub-normals. In the present case, an even more modest goal was established — to provide estimates of the number of subjects in broad categories of verbal intelligence. Therefore, only the verbal subscale was used in an untimed manner, and estimates of "intelligence" were based on this limited subscale. Therefore, the measure used here should not be over-interpreted. It is used only to guarantee that

subjects with lower- or upper-level intelligence were not over-represented in the sample, and to provide some assurance that intelligence was adequately sampled within the age groups.

**The FAA-RTP Test**

For this effort, an FAA-specific version of the NovaScan test paradigm was created (O'Donnell, 1992; O'Donnell, 1993a; O'Donnell, 1993b). NovaScan is a test framework in which specific tests can be inserted, depending on the application of interest. Generically, it can be described as a multitasking situation that controls stimulus sequencing with some degree of precision. It is designed to be administered as a brief test (3 to 10 minutes) that probes a number of separate cognitive and performance functions in the individual.

In the present case, the test developed for the FAA was geared toward the types of functions required by Air Traffic Control Specialists (ATCSs). This was done through a broad literature survey (e.g., Computer Technology Associates, 1987; Redding, Cannon, Lierman, Ryder, Seamaster, and Purcell, 1990; Rodgers and Drechsler, 1993; Seamaster, Redding, Cannon, Ryder, and Purcell, 1992; and unpublished FAA documents) of the types of skills typically utilized by ATCSs in their normal environment. In addition, the level of attention and task multiplexing required by a variety of air traffic control jobs were analyzed. These analyses revealed that common (though variable) elements of these jobs included spatial visualization and situational awareness (generically referred to as "seeing traffic"), vigilance skills (detecting events), and remembering both verbal and spatial information while carrying out other distracting tasks. Based on this, an FAA-specific version of the NovaScan test was designed. This FAA-RTP test and data produced by it, are described below.

The FAA-RTP test consists of 3 separate tasks required of the subject: 1) the "vector" task, 2) the "matrix" task, and 3) the "angles" task (Figure 1). The subject sees 1 of the first 2 of these (i.e., either the vector or the matrix task) in the center of the computer screen, and is required to perform that task. The third task (angles) is always on the screen, and must be performed concomitantly with the other one on the screen. The vector and matrix tasks alternate in an apparently random fashion, so that the subject must continuously switch from doing one task to doing the other. The angles task requires the subject to make a response whenever a certain configuration appears. The subject is instructed that this task takes precedence over either of the other

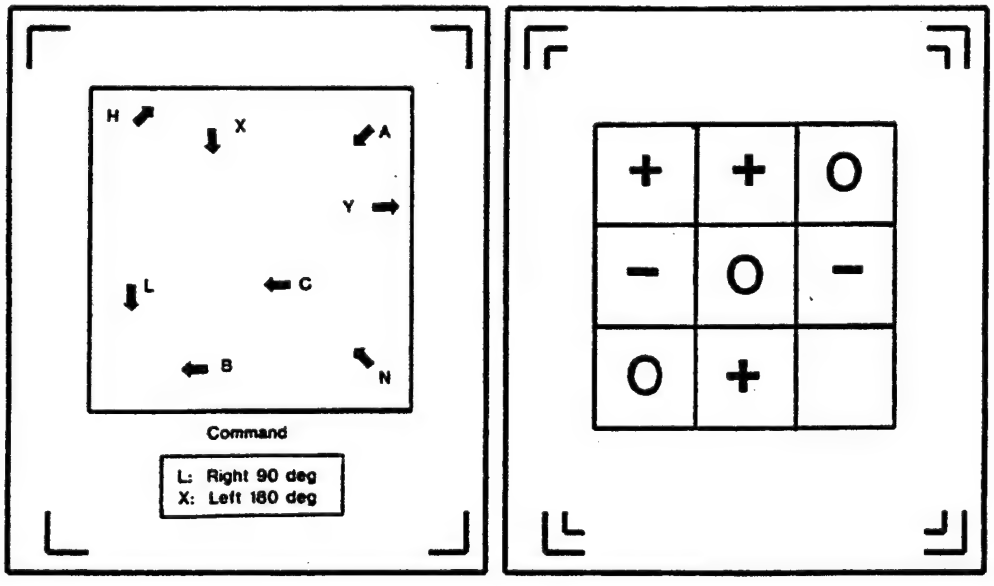



**Figure 1.** Representative Screens of the FAA-RTP Test.

tasks. In the present experiment, a total of 40 trials on each type of task was presented in the course of a given FAA-RTP test.

**The Vector Task**. The subject sees 8 small arrows on the computer screen. Each arrow is pointed in a direction (limited to 0, 45, 90, 135, 180, 225, 270, or 315 degrees on the compass), and is designated by a single letter of the alphabet. Below the display, a "command" line tells the subject that 2 of the arrows "want" to turn by a given amount and in a given direction (e.g., "E: 90 degrees left and A: 180 degrees right"). These "turns" are limited to 90-degree increments. The subject's task is to visualize what would happen if both of the arrows turned in the "requested" direction. Projecting the pathway of the "turned" arrows to infinity (and ignoring all the other arrows), would they ever intersect or "hit" each other? If so, one response is given; if not, another response is given. In summary, the subject's task was to look at the command line to note which arrows were the targets and which turns they wanted to make, and then to (1) visually search for the designated arrows, (2) perform the requested mental rotations, and finally, (3) decide whether the projecting path of the arrows after their resulting rotations would ever cause them to cross, or "conflict."

In actual implementation, target arrows could appear only in non-adjacent-cell positions, thus producing 78 possible pair locations. Given the possibility of eight orientations, there are 11,232 total possible combinations. Eliminating positions of the target arrows that would either be extremely easy (two 90-degree turns of already non-intersecting arrows) or extremely difficult (arrows in the extreme corners of the matrix) produced a more manageable selection of the remaining pairs (96 conflict pairs and 96 non-conflict pairs). These 192 pairs were randomized and selected from the random-order list for each trial, and presented in accordance with the criteria listed below.

**The Matrix Task**. The subject is shown a 3 X 3 matrix, filled in with 8 symbols. The possible symbols are "+", "-", and "o". If there were 3 of each symbol, all 9 cells of the 3 X 3 matrix would be filled. However, one symbol is missing (1 cell is blank). The subject's task is to note **which symbol is missing, and where the missing cell is located**. On the **next** presentation of the matrix, the subject must decide whether the same symbol is missing from the same position as in the previous matrix. If so, the response is "same" — if not, the response is "different." Thus, there can be 2 ways a "different" response can be required: 1) the symbol missing in the first presentation can be present in the second, or 2) the missing symbol can be the same, but the empty cell is in a different position in the matrix. Further, the subject must often retain the spatial (and/or verbal) representation of the first stimulus configuration while responding to the vector test described above for an indefinite number of trials.

**The Angles Task**. The third task required of the subject is a simple "monitoring" task that is intended to simulate the common need in air traffic control to detect a relatively "rare" event which, while critical, is peripheral to the prime duty. During presentation of each of the 2 tasks described above, the subject sees a "frame" surrounding the vector or matrix screen. This rectangular frame is defined by "angles" at each of the 4 corners of the frame. These angles consist of either 1 or 2 carets. The subject's task is simply to determine whether all of the carets are singular, or all of the carets are double. If dual carets are detected, a response is required (i.e., if an "unusual" situation exists, it must be attended to). If all carets are singular, no response is required.

The 3 tasks of the FAA-RTP procedure are counterbalanced and controlled so that each presentation of the FAA-RTP test, while appearing random to the subject, is equated according to the following criteria:

1) The vector task trials contained half conflict, and half non-conflict arrow pairs, per given FAA-RTP test.

2) The number of times that either the vector task or the matrix task was presented sequentially was not permitted to exceed 4. In other words, neither task appeared more than 4 times in a row.

3) The number of times that each of the above tasks appeared sequentially was counterbalanced. In other words, the vector task appeared 1 time prior to a matrix task trial an equal number of times as it appeared 2, 3, or 4 times prior to a matrix task trial.

4) For the vector task, the number of 90- and 180-degree rotation commands were equal.

5) Trials were balanced in the vector task such that an equal number of conflict and non-conflict pairs of arrows appeared following transitions (i.e., immediately following the matrix task).

6) Trials were balanced in the matrix task such that an equal number of same and different responses were required following transitions (immediately following the vector task).

7) Trials were balanced in the angles task such that a response was required (carets changed from single to double) an equal number of times for both transition and non-transition trials.

For each test administration, the subject was first given a "warm-up" session consisting of 24 trials. The purpose of the warm-up was, of course, to assure that the subject correctly recalled the procedures and was otherwise prepared for the test. During training, the warm-up also provided feedback to the subject in the form of scores, which provided summary statistics on the reaction time, standard deviation, and percent correct for the vector and the matrix tasks and data on the number of angle tasks ("acknowledgments") presented and requiring detection. A sample of the warm-up feedback screen seen by the subject is shown in Figure 2.

**Apparatus**

***FAA-RTP***

The FAA-RTP test utilizes a basic 286 or higher IBM- compatible computer, along with a specialized response unit. In the present experiment, the standard FAA "OATS" computer was used, modified slightly, as described below, to accept the

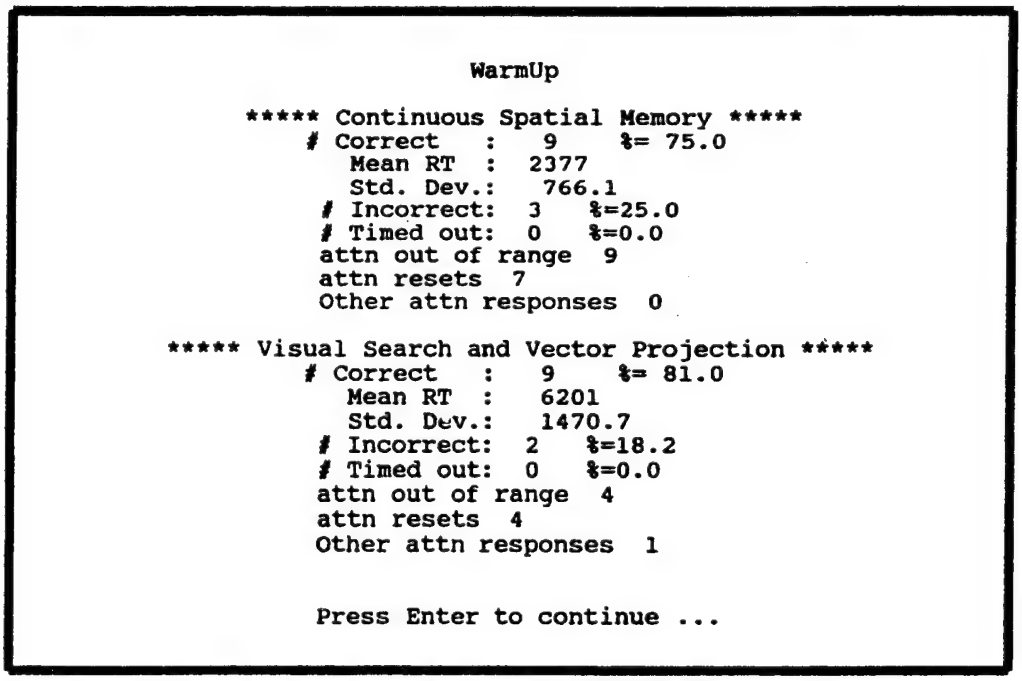



**Figure 2**. Sample Warm-Up Results Screen.

unique NovaScan response unit. This computer system is an AT&T-supplied system, consisting of a 286 processor, with 300 mb hard drive, and one 3 1/2-inch floppy drive. The basic system is configured for a 9.6 kbaud rate. However, the FAA-RTP response unit requires a 19.2 kbaud rate. Thus, new serial ports were installed into the OATS systems for this experiment.

The NovaScan response units (Model RU-1) were supplied by Nova Technology, Incorporated, Calabasas, California, for this experiment. This unit (Figure 3) provides 6 response buttons (3 for each hand) plus a keypad. A joystick is also included, but was not used in this experiment. As shown in Figure 3, the subject's left hand is positioned over the left 3 response keys, and the right hand is positioned over the right 3 response keys. These positions were standardized for all subjects in the experiment. The RU-1 provides near-millisecond timing (less than 3 milliseconds error per response) when associated with the FAA-RTP test software.

***Breath Alcohol Concentration***

Blood alcohol concentration was estimated using the Intoxilyzer 5000 produced by CMI, Inc., which measures breath alcohol content (BrAC) in terms of grams per 210 liters of air. This unit was calibrated by the manufacturer prior to the study. FAA personnel checked the calibration of the system according to the manufacturer's instructions to within plus or minus .001 percentage point. For example, tolerances of between .039 and .041, in order to achieve .040 percent BrAC, would be acceptable. The Intoxilyzer 5000 automatically induces a "clearing" period of 90 seconds between breathalyzer tests, and in the present experiment, at least 90 seconds were allowed between tests to assure that readings would not be contaminated.

The actual procedures for utilization of the intoxilyzer included having each subject use a personal mouthpiece. This mouthpiece captures particulate matter and assures that contamination from one reading to another is eliminated. Subjects inserted

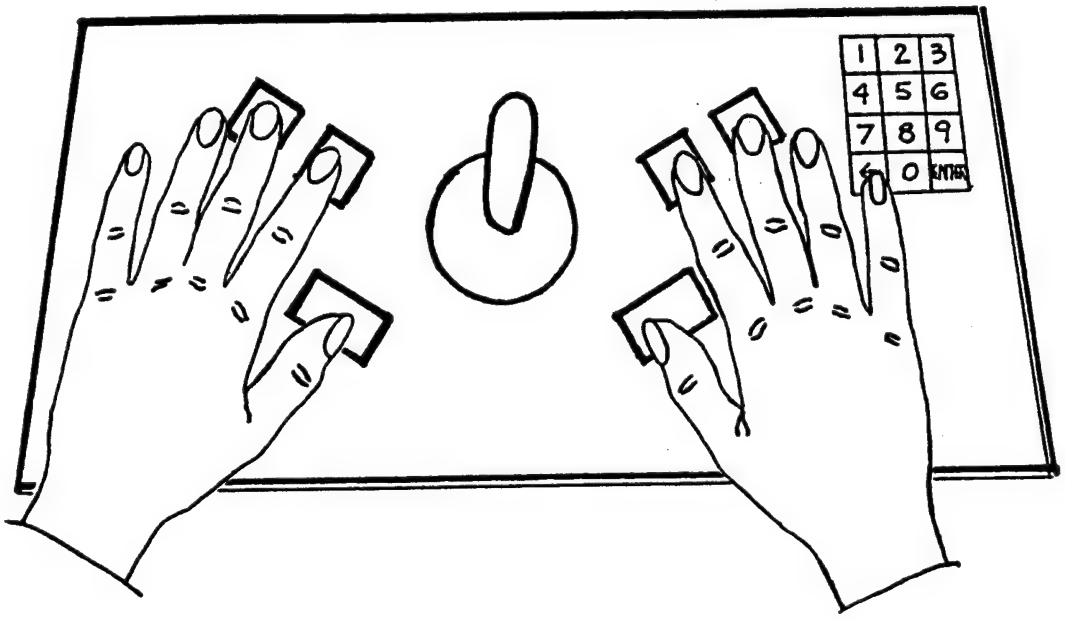



**Figure 3.** FAA-RTP RU-1 Response Unit.

the mouthpiece into a tube and began to expire air forcefully. An auditory tone indicated when the sample was begun, and the subject blew into the tube as long as the tone remained on. When an adequate sample had been collected, the tone stopped and a final reading was given to the technician by means of an LED display. The subject, of course, could not see this reading. Identical procedures were followed in obtaining BrAC readings on both the alcohol and placebo days. BrAC determinations were made before and after performance on the FAA-RTP test, and the reported level used in this report is the average of these 2 determinations.

**Training Methods**

The entire subject group was arbitrarily split into 2 sub-groups for training and testing. The first subgroup was trained during the first week of the experiment, and tested during the second week. The second sub-group was trained during the third week and tested during the fourth week.

Prior to the experiment, it had been estimated, based on previous experience with other NovaScan procedures, that 12 to 15 practice sessions on the FAA-RTP test would be sufficient to establish stable baselines. Considering additional time necessary for administrative details, initial explanations, and training on the Mah Jongg game, subjects in the first group were scheduled for 8 hours of training time, distributed over 4 sessions in 2-hour sittings. The first hour of the first 2-hour period was spent in a general introduction to FAA-RTP and to the entire project. The topics covered during this hour included the following: 1) brief history of NTI's affiliation with the FAA, 2) purpose of the program, 3) brief history of performance-based testing and NovaScan, 4) review of time-card procedures, 5) review of protocol and subject voluntary consent agreement, 6) overview of testing schedule and travel arrangements, 7) instructions and completion of medical and alcohol questionnaires, and 8) instructions on NovaScan and Mah Jongg. The Subject Voluntary Consent Agreement (NTI, Inc., 1993) was read aloud while subjects read along, and every opportunity was given for individuals to ask questions or to withdraw from the study.

Following this initial introduction, subjects began practicing on the FAA-RTP and Mah Jongg tasks described above. Such practice consisted first of a broad tutorial on the FAA-RTP tasks. This tutorial was accompanied by a handout booklet explaining the tasks and illustrating the various screens and apparatus to be used (see Appendix A).

After all questions were answered, subjects began hands-on practice with the FAA-RTP tests. As required, experimenters assisted each subject individually to understand the procedures and to adopt appropriate strategies in handling the test. As soon as subjects were able to perform the tests independently, the experimenter permitted ad lib practice for the duration of the practice time, making only occasional comments, suggestions, or encouragements for individual subjects.

In addition to the feedback provided by the experimenter during the training session, each subject was provided with a "feedback data sheet" (ibid.), and was instructed to record on that sheet the warm-up scores obtained for each FAA-RTP test. In this way, the subject was able to observe his own level of proficiency as it improved over practice sessions. This provided some additional feedback to the subject for training purposes, and also permitted the experimenter to detect subjects who were having particular problems and provide additional instructions to those subjects.

During the end of the first week of training, it became evident to the experimenters that a small number of subjects were continuing to have difficulty with the FAA-RTP test. Careful inspection of the warm-up feedback sheets indicated that, while many subjects had reached plateau rapidly, a small number had still not progressed satisfactorily during the 8 hours of practice. Therefore, a decision was made to provide all subjects with an additional 2 hours of practice. This 2-hour practice session was again carried out on an ad lib basis, with intensified instruction and corrections by the experimenter for those subjects appearing to have difficulties.

Subsequent to this additional training, the learning curves actually produced by this first group of subjects over the 10 total hours (including training

on both NovaScan and Mah Jongg) were visually inspected by the principal investigator to determine whether they appeared to reach plateau. It was concluded that the majority had produced apparently stable baselines, but that some subjects baseline stability was still highly questionable. The most questionable subjects were therefore requested to appear for yet another 2 hours of training. Based on this additional training, these subjects were subsequently run on tests days. However, their data were inspected after completion of the study to determine whether they should be included in the final analysis.

Because the original estimate of training time was apparently too short for a small number of subjects, the training regimen for the second group of subjects was changed slightly. Each subject in this second group was given a total of 11 hours of training, distributed over 4 days, with 2 hours and 45 minutes per day of practice. Visual inspection of the learning curves suggested that this training time appeared sufficient for most of the subjects in the second group. Additionally, a third group of subjects was recruited, then trained following the regimen described for the second group, once it was realized that attrition had significantly reduced the N for the younger group of subjects.

The above total training times for each of the 3 groups also included training on the Mah Jongg task. Subjects practiced the Mah Jongg game for a total of 15 to 25 minutes per test session (i.e., 4 to 5 times over the training week). Typically, the game was introduced after 1 hour of FAA-RTP practice. Subjects were given a brief introduction to the Mah Jongg game, and the basic rules were explained. They were then allowed to practice this game as many times as they could during the allotted time. To assure that subjects would be motivated to perform the Mah Jongg game as efficiently as possible, they were informed that a $20.00 bonus was available for performing well on this task. In fact, every subject who completed both test days of the experiment was given the $20.00 bonus.

**Detailed Procedures**

Subject testing followed the schedule outlined in Table 1. On each test day, subjects were picked up at home by Government drivers, who took them to the Civil Aeromedical Institute building. Subjects usually arrived at the building between 10:30 and 11:30 a.m., and were gathered together in a common room where they were permitted to watch television and to eat a light lunch. Subjects had been instructed to eat a normal breakfast and to bring a light lunch with them. It was recommended that the lunch not have fatty items (e.g., peanut butter), but rather consist of a simple sandwich and some fruit. However, no control was exercised over this food intake.

At 11:30 a.m., subjects were given a brief introduction to the procedures to be followed during that day, including instructions on the use of the breathalyzer. They then immediately proceeded to the breathalyzer machine, where the baseline breathalyzer test was administered. Following this, subjects proceeded to the testing room and took the first baseline (FAA-RTP test) and filled out the first fatigue questionnaire. After this, they reported to the "drinking" room for the first drinking session.

At approximately 12:10 p.m., subjects began the first drinking session, designed to raise their BrAC to .02%. A designated "drinking room," was used for this purpose was equipped with a television set and reading material. The technician provided the subjects with an orange juice drink, which might or might not contain a significant amount of vodka, depending on their assigned experimental condition. On the "non-drinking" day, a placebo drink was prepared, consisting of 5 mls. of vodka mixed with the same amount of orange juice used on the "drinking" day.

Subsequent to the 10-minute drinking period, subjects reported back to the testing room, and began performing the Mah Jongg task. At approximately 12:35 p.m., subjects reported for their second breathalyzer test. This was again followed by the FAA-RTP test and a post-FAA-RTP breathalyzer

**TABLE 1**

APPROXIMATE TIMES FOR EACH EVENT DURING TEST DAYS

Morning

Subjects instructed to eat a normal breakfast and to bring a moderately light lunch to be consumed prior to 11:30 a.m.

| | |
|---|---|
| 11:30 | Subjects assembled and instructed — Measure and record BrAC, rate fatigue, FAA-RTP warm-up and baseline. |
| 12:10-12:25 | .02% drinking session. Mah Jongg game performed throughout day at all times when other interfering activities were not scheduled. |
| 12:35 | Measure and record BrAC, rate fatigue. |
| 12:45 | FAA-RTP and BrAC measurement. |
| 13:15-13:30 | .04% drinking session. |
| 13:40 | Measure and record BrAC, rate fatigue. |
| 13:45 | FAA-RTP and BrAC measurement. 10 min. break. |
| 14:15-14:30 | .06% drinking session. |
| 14:40 | Measure and record BrAC, rate fatigue. |
| 14:45 | FAA-RTP and BrAC measurement. 10 min. break. |
| 15:15-15:30 | .08% drinking session. |
| 15:40 | Measure and record BrAC, rate fatigue. |
| 15:45 | FAA-RTP and BrAC measurement. 10 min. break — consume food. |
| 16:25 | Measure and record BrAC, rate fatigue. |
| 16:30 | FAA-RTP and BrAC measurement. |
| 17:25 | Measure and record BrAC, rate fatigue. |
| 17:30 | FAA-RTP and BrAC measurement. 10 min. break. |
| 18:25 | Measure and record BrAC, rate fatigue. |
| 18:30 | FAA-RTP and BrAC measurement. 10 min. break. |
| 19:25 | Measure and record BrAC, rate fatigue. |
| 19:30 | FAA-RTP and BrAC measurement. |
| 19:45 | Measure and record BrAC. Subjects leave in Government-provided transportation. |

test. This procedure was repeated with all subjects at approximately 1 hour intervals until approximately 3:15 p.m., when the subjects received their last drink. At that point, subjects were permitted to intake their first food since 11:30 a.m. This consisted of ad lib amounts of dry, lightly salted crackers, apples, and bananas. Twenty minutes were allotted for this food intake, after which the cycle of breathalyzer, fatigue scale, FAA-RTP, breathalyzer, and Mah Jongg was carried out. This, again, was done on an approximate 1-hour basis until 8:45 p.m., when the test run was terminated. At that time, subjects were requested to fill out the post-test-day questionnaire. They were then driven to their home by Government transportation.

The formula for titrating BrAC in each subject was based on a procedure adapted from a previous study conducted by the Civil Aeromedical Institute's Human Factors Laboratory (Schroeder, et al, manuscript in development). This procedure generally

followed the recommendations of Lentz and Rundell (1976). On the alcohol day, the first drink consisted of 225 mls of orange juice mixed with .19 mls of 80-proof vodka per kg of body weight. The second drink consisted of 100 mls of orange juice mixed with .14 mls of 80-proof vodka per kg of body weight. The third and fourth drinks consisted of 100 mls of orange juice mixed with .16 mls of 80-proof vodka per kg of body weight. All drinks contained 2 oz of crushed ice. Drinks were consumed over a 10-minute period. The second through fourth drinks were adjusted (plus or minus .08 mls per kg of body weight) for some of the subjects, based on their BrAC during the previous testing session. Using this formula, most subjects achieved the desired BrAC at the desired time (plus or minus .01% BrAC). However, a few subjects failed to achieve the targeted BrAC levels with the above procedure, and these subjects were given a "booster" drink of .034 mls per kg of body weight in a 1:4 mixture of orange juice. When this occurred during the first test day, a "sham" booster drink was also administered to the subject at the same time during the placebo day.

## SECTION 3
## RESULTS

**Overview**

This section is divided into 3 separate subsections, covering the various measurements and questions of interest in the present study. In the first section, the subject population is described in terms of age distribution, verbal "intelligence" level, occupation, and drinking history. The second subsection deals with the various FAA-RTP measures, and the effect of experimental manipulations on those measures. This subsection is divided into separate discussions of training results and learning curves, measures of fatigue, and a large section on FAA-RTP results under alcohol conditions. In this latter section, a further subdivision is made between statistical assessment of FAA-RTP group effects, and analysis of individual FAA-RTP performance scores and the predictability of the tests, with respect to a breath alcohol criterion. Finally, data on ancillary questions of interest are presented in the third section. These questions involve primarily the time to take the FAA-RTP test, and whether this time could be reduced.

**TABLE 2**

AGE GROUPS OF PARTICIPATING SUBJECTS

| Age Group | N | Mean | Median |
|---|---|---|---|
| 25-34 | 25 | 29.72 | 29.25 |
| 40-48 | 27 | 43.37 | 43.14 |
| 54-62 | 25 | 58.28 | 57.64 |

**Description of the Subject Population**

*Age*

Three target age groups were established for this study: 25-33, 40-48, and 54-62. Subjects were recruited within these age categories, as described above. The final distribution of ages for subjects completing the study is shown in Table 2 below. One subject in the older age group had turned 63 between being recruited for the study and actual testing. Thus, the actual range for this group was 54-63. However, for simplicity, this group will continue to be called the "45-62" group.

Table 2 reveals that the ages were nearly evenly distributed within each of the samples. The overall average age was 43.71, and the median age was 38.13 for all 77 subjects.

***"Intelligence"***

Based on the results of the Shipley Institute of Living Scale (Shipley, 1940), the WAIS equivalent IQ estimates (based on tables provided by Crumpton, et al, 1985) are presented in Table 3 as a function of age group. It can be seen that no subject scored below an equivalent IQ of 83, with the group means for each age group being 105, 106, and 114, respectively, for the younger, middle, and older age groups. Overall mean IQ equivalent for the subject population was 108.

It appears from these results that the older age group was somewhat superior in verbal IQ to the others. The Shipley scale was not used as a timed test in this case, and therefore would not be expected to be unduly influenced by age. Further, there might have been a selection factor operating in that some of the older subjects were individuals who had retired from demanding jobs, while some of the younger subjects were unemployed laborers. This effect showed itself in a higher percentage of older subjects scoring in the 120-129 IQ range, and a lower percentage in the 80-109 IQ range. Since there was only a 12% range in subjects between 100 and 119, however (and these accounted for the majority of subjects), it would appear that "intelligence" should not be a causal factor accounting for any age differences that might be observed in other variables. In future studies, however, it would be desirable to match the intellectual level of subjects more precisely to targeted FAA populations.

**TABLE 3**
WAIS EQUIVALENT IQ ESTIMATES BY AGE GROUP

| | AGE GROUP | | | |
|---|---|---|---|---|
| | 25-34 | 40-48 | 54-62 | All |
| VERBAL IQ ESTIMATE | | | | |
| 120 -129 | 0 (0%) | 0 (0%) | 7 (28%) | 7 (9%) |
| 110 -119 | 11 (44%) | 10 (37%) | 11 (44%) | 32 (41%) |
| 100 -109 | 9 (36%) | 12 (44%) | 6 (24%) | 27 (35%) |
| 90 - 99 | 4 (16%) | 4 (15%) | 1 ( 4%) | 9 (12%) |
| 80 - 89 | 1 ( 4%) | 1 ( 4%) | 0 ( 0%) | 2 ( 3%) |
| N | 25 | 27 | 25 | 77 |
| Mean | 105.48 | 105.81 | 114.36 | 108.48 |

**TABLE 4**

AGE GROUP AND SELF-REPORTED ALCOHOLIC PRACTICES
BASED ON REPORTED FREQUENCY OF DRINKING

| Drinking Category | 25-34 | 40-48 | 54-62 | TOTAL |
|---|---|---|---|---|
| "Heavy" Drinker | 11 | 3 | 3 | 17 |
| "Moderate" Drinker | 6 | 7 | 5 | 18 |
| "Light" | 4 | 13 | 9 | 26 |
| "Infrequent" | 3 | 2 | 4 | 9 |
| "Abstainer" | 1 | 2 | 4 | 7 |
| Totals (N) | 25 | 27 | 25 | 77 |

***Drinking History***

A self-report of the subjects' drinking history was obtained from a modified version of the alcohol history form used by the Alcohol Research Unit at the University of Oklahoma (NTI, Inc., 1993). The principal question on this form concerns the frequency of the subject's drinking. Such amounts are divided into nine categories, ranging from "more than 3 drinks per day" to "no beverages containing alcohol in last 6 months." Individuals were grouped into heavy, moderate, light, infrequent, or abstaining drinkers based on criteria established by the Quantity-Frequency-Variability Index (Cahalan, Casin, and Crossley, 1967). The breakdown of the subject population by age, according to this classification, is shown in Table 4.

Although self-reports of alcohol consumption are frequently low and unrealistic, this table still reveals an over-representation of "heavy" drinkers among the younger age group (44% vs. 11% and 12% for the other age groups). Conversely, 48% of the middle age group and 36% of the older group were classified as "light" drinkers, as compared to 16% of the younger group in this category. This breakdown may reflect drinking patterns in the general population, but is more likely due to the socio-economic pattern of the particular sample tested in this experiment. As noted earlier, the older group tended to have a slightly higher intelligence level, and probably included more retired, high-achieving individuals than the younger group. This latter group probably contained a significant number of unemployed individuals with current problems. In view of this, it is especially important that the results of the present study be interpreted with great caution, as they might apply to a population of FAA individuals in safety-sensitive positions.

**BrAC Levels Achieved During the Test**

The formula used to titrate the BrAC of subjects proved to be extremely accurate as it was applied by the FAA technicians. BrAC measures in individual subjects seldom varied by more than .01% from the targeted figure during the ascending limb of the BrAC curve. During the descending limb, of course, more variability was seen, due to the variations in metabolism among the subjects. The mean and variability measures in BrAC for each testing period are shown in Table 5. It can be seen that, on average, the targeted levels were achieved (±.01%). In fact, the raw data revealed that the targeted levels were hit **precisely** in 80.52% of the cases on the ascending limb. For the descending limb, the test periods yielded average BrAC values of .055, .039, .022, and .009 percent, respectively. These mean values will be used in subsequent tables of this report. However, it should be remembered that the increased variability of these latter values means that precise statements about BrAC-test relationships must be interpreted with some caution.

**TABLE 5**

AVERAGE BREATH ALCOHOL CONCENTRATION AT EACH TEST PERIOD

| BrAC TEST | TARGET | N | MEAN BrAC | SD |
|---|---|---|---|---|
| 1 | .02% | 77 | .021 | .005 |
| 2 | .04% | 77 | .039 | .005 |
| 3 | .06% | 77 | .061 | .005 |
| 4 | .08% | 77 | .079 | .007 |
| 5 | | 77 | .055 | .013 |
| 6 | | 77 | .039 | .014 |
| 7 | | 77 | .022 | .015 |
| 8 | | 77 | .009 | .011 |

**FAA-RTP Training Results**

Since the FAA-RTP test represents a new implementation of the NovaScan paradigm, there was considerable interest in determining the precise amount of training necessary for test subjects to reach plateau levels. Therefore, an extensive analysis of the training data was carried out. Training procedures have been described earlier, as well as the fact that different groups of subjects received different amounts and schedules of training. However, all subjects received distributed practice for at least several hours of training.

The FAA-RTP test is capable of yielding a large number of scored variables (approximately 160) based on the types of analyses carried out on each of the separate tests. Although any of these variables potentially could be used as dependent measures in the present study, many are inverses of the other (e.g., "percent correct" and "percent incorrect"). In addition, past experience with other implementations of the NovaScan procedure revealed that a small sub-set of the possible measures turns out to be significant in most studies (O'Donnell, 1993a). Essentially, total reaction times, transition reaction times, and the percent correct measures associated with them have proven to be the most sensitive. In some studies, a version of the "thruput" measure described by Thorne, Genser, Sing, and Hegge (1985) has also proven sensitive.

Based on these earlier results, 15 "simple" variables (not utilizing complex combinations of single variables) were chosen as a first candidate set, and are used for subsequent analyses in this report. The 15 variables chosen, and their abbreviated designators used in the remainder of this report, are shown in Table 6.

**Analysis of Averaged Training Data**

The median subject received a total of 22 training sessions each on FAA-RTP (range from 12 to 44 sessions). Obviously, the reason that different subjects received different amounts of training was that they differed in their base reaction time on the tests, allowing each subject to complete a different number of tests in the time allowed.

Baseline points (the point where learning was stable enough to begin collecting baseline data) for each individual were first subjectively estimated by a single judge. These were based on the usual criterion of when the rapidly accelerating portion of the learning curve had been completed, and when the session-to-session variability was low enough to possibly collect meaningful data. These baseline points ranged from session 6 to session 39 for various subjects, with a median value of 16.5 sessions

**TABLE 6**

CANDIDATE FAA-RTP VARIABLES CHOSEN FOR STUDY

| DESIGNATOR | DESCRIPTION OF VARIABLE |
|---|---|
| T1RT | Task 1 (vector) reaction time |
| T1PC | Task 1 (vector) percent correct |
| T1TRRT | Transition from task 2 to task 1 reaction time |
| T1TRPC | Transition from task 2 to task 1 percent correct |
| T1TRTP | Transition from task 2 to task 1 "thruput"* |
| T1TP | Task 1 (vector) "thruput"* |
| T2RT | Task 2 (matrix) reaction time |
| T2PC | Task 2 (matrix) percent correct |
| T2TRRT | Transition from task 1 to task 2 reaction time |
| T2TRPC | Transition from task 1 to task 2 percent correct |
| T2TRTP | Transition from task 1 to task 2 "thruput"* |
| T2TP | Task 2 (matrix) "thruput"* |
| ATTNRT | Angles task (attention) reaction time |
| ATTNPC | Angles task (attention) percent correct |
| ATTNTP | Angles task (attention) "thruput"* |

* "Thruput" is defined as the reaction time divided by the proportion of correct responses. It is not to be confused with the "thruput" term as traditionally used in communication theory. Rather, it is a simple correction for the number of incorrect responses.

(average 16.4 sessions). The 10th and 90th percentile number of sessions were 9.1 and 23.4. In other words, if subjects took an average of 10 minutes to complete early training sessions (including rest periods), a total of 2.75 hours of training would have been required for 50 percent of the subjects to reach the point where meaningful "baseline" data could start to be collected on the FAA-RTP test. To reach the point where 90 percent of the subjects would have reached baseline, a total of 3.9 hours of training would be required.

In order to obtain an estimate of the averaged "pattern" of skill acquisition on this task (as opposed to actual time), a unique type of "learning curve" was calculated for each variable. Since subjects received different amounts of training to achieve plateau or baseline levels of performance, it is not legitimate to average subjects' data together on a session-by-session basis. Such a procedure would result in a session from a subject who had already reached plateau being averaged with one who was still learning rapidly, and would distort the shape of the curve. Interest here is in determining the overall pattern of skill acquisition (rather than its timing). Therefore, this pattern can be best revealed by dividing the total training time of each subject into a fixed number of "periods," and taking data for each of these periods as the elements of the learning curve.

Since the subjective analysis above suggested that a median of 16.5 sessions was sufficient to reach baseline, it was decided to take 17 sessions as the fixed number. Therefore, each subject's training data were divided into 17 equal periods. Within each period, an "average score" was obtained from all included tests, and this was taken as the score

for that period. In this way, a "learning curve" was constructed for each subject that had 17 data-points on the abscissa. Subjects with fewer than 17 training sessions were eliminated from this analysis. Although several of these subjects appeared to have reached baseline, we felt that inclusion of their data might result in an artificially low estimate of the time needed for training. The resulting basic curves are, therefore, based on 62 subjects. These data are presented in Figure 4 on the next three pages. In these figures, the 17 average values are shown. In addition, the performance of 37 subjects who took the "fatigue" day first are appended to the end of the learning curve. This provides an additional 9 data points (although based on fewer subjects) and allows comparison of how much the subjects may have continued to improve after formal training was terminated.

Objective determination of how well the given learning curve fits classical models is somewhat problematical (Damos, 1991). However, a good criterion appears to be a degree to which the observed curves fit a theoretically interpretable mathematical function (Speers, 1985). Measures of the "loss function" of such a fit provide a clue as to how much deviation there is from the idealized curve. Classically, this is defined as the sum of deviation squares around the predicted value. Further, the terms of the model provide estimates of various parameters of the curve, such as the asymptote, rate of learning, and initial value. This analysis was carried out on the learning curve data generated above, and the theoretical curves are overlaid on the actual data in Figure 4. It was found that both the reaction time and percent correct scores were adequately fit with a classical exponential function of the form:

$$Y = K_1 + \{K_2 * e^{(-K_3 * X)}\}$$

where: e = base of the natural logarithm
$K_1$ = asymptote
$K_2+K_2$ = initial value
$K_3$ = rate of learning

**TABLE 7**

FIT OF THE OBTAINED LEARNING CURVES TO AN EXPONENTIAL MODEL

| FAA-RTP VARIABLE | ASYMPTOTE | INITIAL VALUE | 1-RESIDUAL |
|---|---|---|---|
| T1RT (ms) | 8130 | 13268 | .96 |
| T1PC (%) | 94 | 79 | .88 |
| T1TRRT (ms) | 8573 | 14339 | .95 |
| T1TRPC (%) | 95 | 90 | .51 |
| T1TRTP (ms) | 9245 | 16097 | .96 |
| T1TP(ms) | 8882 | 15266 | .98 |
| T2RT (ms) | 2759 | 4473 | .97 |
| T2PC (%) | 93 | 70 | .83 |
| T2TRRT (ms) | 3372 | 5271 | .95 |
| T2TRPC (%) | 85 | 55 | .64 |
| T2TRTP (ms) | 4327 | 7538 | .95 |
| T2TP (ms) | 3081 | 5515 | .98 |
| ATTNRT (ms) | 755 | 1347 | .97 |
| ATTNPC (%) | 99 | 99 | .19 |
| ATTNTP (ms) | 765 | 1464 | .97 |

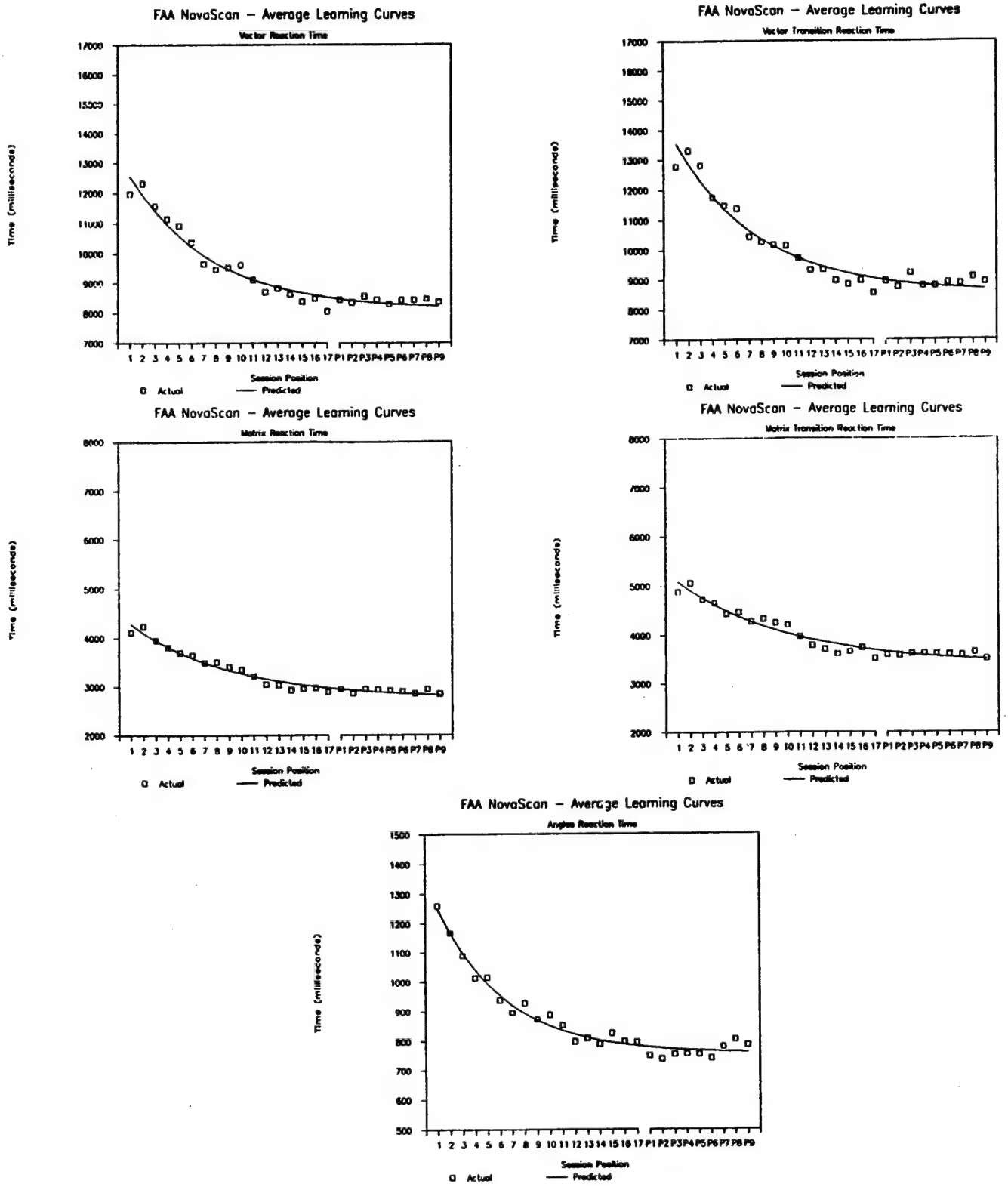



**Figure 4a**. Averaged Learnings Curves.

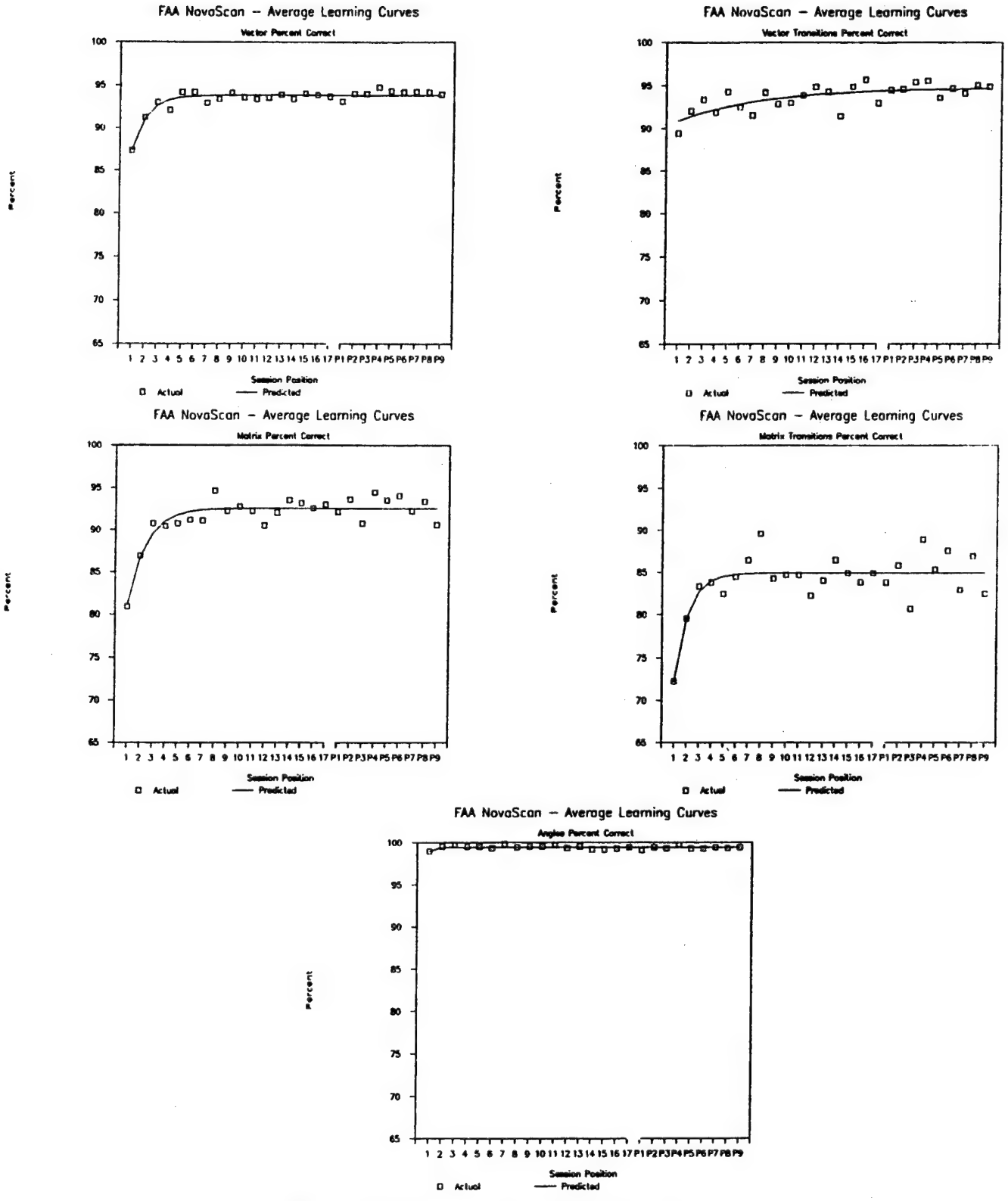



**Figure 4b.** Averaged Learning Curves.

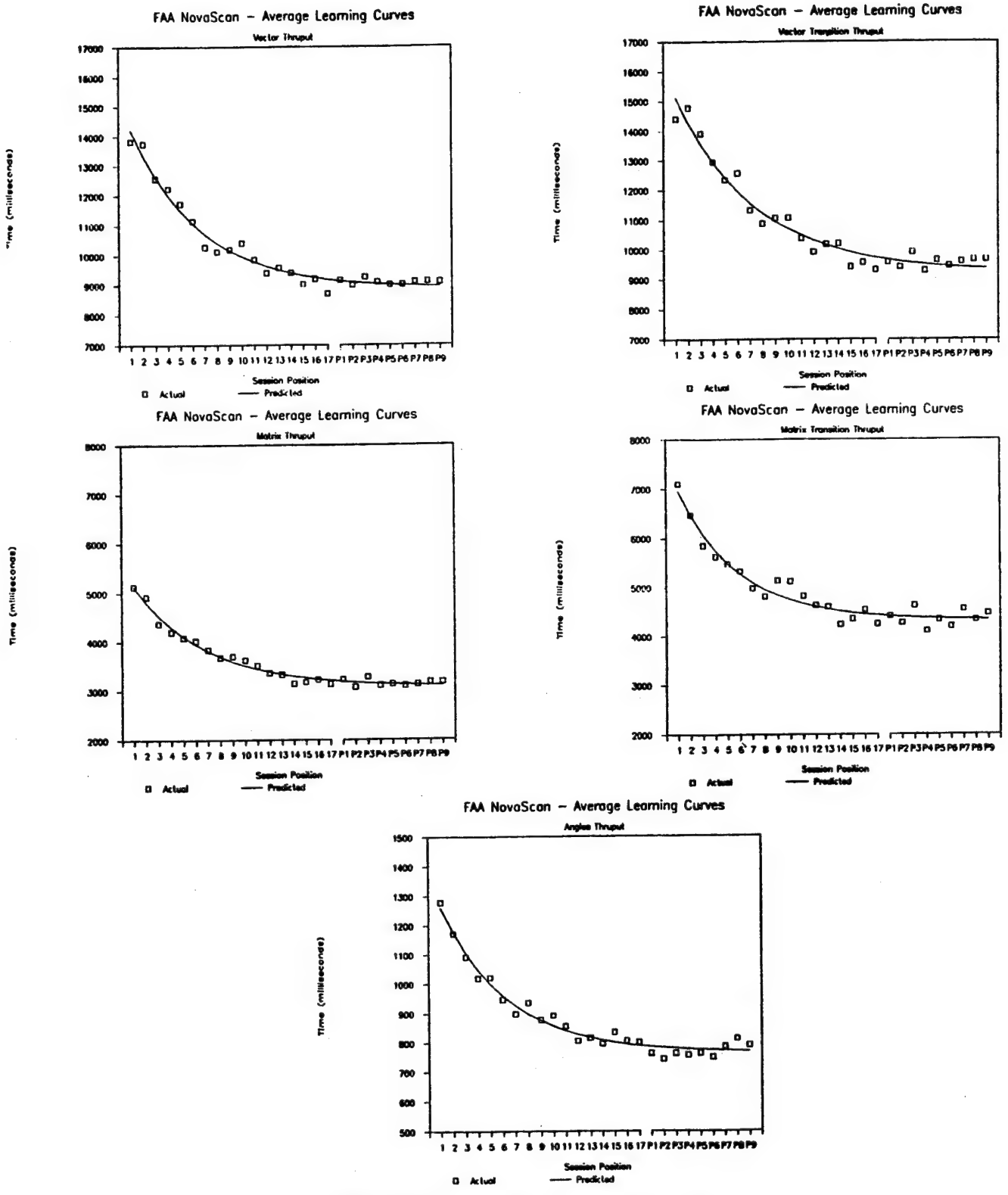



**Figure 4c.** Averaged Learning Curves.

The curves shown in Figure 4 show, for the most part, excellent fit to the predicted exponential function. This is especially true for all reaction time and thruput measures. To quantify the match of the data to the predicted function, the values derived from the curve-fitting procedure are presented in Table 7. In this Table, "asymptote" refers to the predicted plateau level that would be reached by subjects with extensive practice. "Initial value" refers to the starting point of the learning curve, or the subject's "untrained" skill at the task. The residual value is the amount by which the values predicted by the equation differ from the obtained values (the loss function).

First, it appears that the average plateau point for the various reaction time measures in the vector task is between 8 and 9.25 seconds, with a percent correct between 93 and 95%. The matrix task shows a plateau point between 2.7 and 4.3 seconds for the various measures, with accuracy between 85 and 93%. The angles task produced a plateau around .75 seconds, with 99% accuracy. Thus, the vector task appeared to involve the greatest amount of processing. Conversely, the matrix task was easier, but subjects were generally less accurate.

The agreement between predicted and actual values (1-residual) with various FAA-RTP measures suggests that all of the reaction time measures fit the predicted curve with very high accuracy (.95 to .98). As noted above, the percent correct measures appeared to be more variable and not as close to the predicted curves (.19 to .88). The angles task percent correct was especially poor in this respect. It should be remembered, however, that percent correct measures may be unique. Since they show little variation in a statistical sense, they frequently do not show up well in terms of statistical reliability or fit to curves. However, they can still be clinically useful if an occasional individual deviates significantly from normal. Nevertheless, the poor fit to the learning curve demands caution in interpretation of percent correct results.

In general, these figures reveal that reasonably classic patterns of acquisition exist for the majority of FAA-RTP measures, with generally excellent fit to an exponential function. It is not absolutely essential that a learning curve reach plateau in order to be practically useful, since reliability and differential stability have a more important effect on its sensitivity (Jones, Kennedy, and Bittner, 1981). However, it does appear that, at least with respect to group curves for the testing durations used here, the FAA-RTP test displays a stable plateau.

Another way to estimate reliability in this type of test is to calculate the between-subject variance over a set of baseline data, and relate it to the within-subject variance for that same set of data. This analysis, which was suggested by T. Warm of the FAA (personal communication, based on Lord and Novick, 1968), defines reliability ($r_{xx}$) as equal to true score variance (VAR T) divided by total variance (VAR X):

$$r_{xx} = VAR(T)/VAR(X)$$

By a generally accepted ANOVA identity, total variance (VAR X) is made up of true score variance (VAR T) plus error variance (VAR E):

$$VAR(X) = VAR(T) + VAR(E)$$

which converts to:

$$VAR(T) = VAR(X) - VAR(E)$$

incorporating this into the formula for $r_{xx}$:

$$r_{xx} = (VAR(X) - VAR(E))/VAR\ X$$

or:

$$\text{reliability} = \frac{\text{TOTAL VARIANCE - ERROR VARIANCE}}{\text{TOTAL VARIANCE}}$$

In practice, the total variance is estimated by calculating the variance between all scores for all subjects across asymptotic sessions (between subject variance), and the error variance is estimated by calculating the averaged within-subject variance. Thus, the above formula can be expressed:

$$\text{reliability} = \frac{\text{BETWEEN VAR. - WITHIN VAR.}}{\text{BETWEEN VAR.}}$$

This provides an estimate of how stable the individual subject is with regard to the rest of the subject sample. The "reliabilities" calculated in this way for each of the 15 FAA-RTP variables are shown in Table 8.

This analysis revealed that the reaction time measures for all 3 FAA-RTP tests achieved "reliabilities" between .76 and .94. Although these values are not directly comparable to reliabilities calculated in traditional ways, they indicate that individual subjects were showing considerable consistency with respect to the total subject sample. The percent correct measures, on the other hand, that produced reliabilities which were not consistently as high. This appeared to be due primarily to the restricted variability range of the measures. The two percent correct measures, which showed the worst reliabilities (T1TRPC and ATTNPC), each had less than 5 percentage points variability between their initial value and their asymptote (see Table 7). These produced reliabilities of -.07 and .20. The other three percent correct measures yielded reliabilities between .52 and .67. The range for these measures (between initial value and asymptote) was between 15 and 30 percentage points. It, therefore, appears that the low reliability figures for the percent correct measures may reflect that there was not enough variability to permit the statistic to operate properly.

**Analysis of Individual Learning Curves**

The above analysis of average learning curves, while extremely important and informative, could be misleading. Since the ultimate application of the FAA NovaScan test is to make predictions on an individual basis, it is important to know whether there are individual exceptions to the averaged learning curve picture, and if so, where they occur.

**TABLE 8**

ESTIMATED RELIABILITIES OF FAA-RTP VARIABLES

| FAA-RTP VARIABLE | BETWEEN-S VARIANCE | WITHIN-S VARIANCE | "RELIABILITY" |
|---|---|---|---|
| T1RT | 6730372 | 429692 | .94 |
| T1PC | 44.76 | 14.65 | .67 |
| T1TRRT | 8130609 | 844686 | .90 |
| T1TRPC | 43.34 | 34.77 | .20 |
| T1TRTP | 14517224 | 2601419 | .82 |
| T1TP | 12074298 | 842031 | .83 |
| T2RT | 1565657 | 105965 | .93 |
| T2PC | 34.93 | 15.43 | .59 |
| T2TRRT | 2379739 | 285359 | .88 |
| T2TRPC | 123.90 | 59.08 | .52 |
| T2TRTP | 3958270 | 943214 | .76 |
| T2TP | 1898015 | 161150 | .92 |
| ATTNRT | 103979 | 6760 | .94 |
| ATTNPC | 2.88 | 3.08 | -.07 |
| ATTNTP | 112567 | 8224 | .93 |

**TABLE 9**

ASSESSMENT OF LEARNING CURVE QUALITY BY AGE GROUP

| AGE GROUP | QUALITY OF THE LEARNING CURVE VERY GOOD | GOOD | MEDIUM | POOR |
|---|---|---|---|---|
| 25 - 33 YRS. | 11 | 6 | 4 | 3 |
| 40 - 47 YRS. | 14 | 4 | 3 | 4 |
| 55 - 62 YRS. | 8 | 11 | 4 | 2 |
| TOTALS | 33 | 21 | 11 | 9 |

Two subjects out of 77 failed to show any discernible learning curve on the FAA-RTP test. In both cases, there did not appear to be any consistency in their session-to-session performance. It appeared that these subjects were simply not trying to learn, rather than that there was an inability to learn. One subject, 31 years old, had been given 29 training sessions. The other subject had 20 training sessions, and was 60 years old. In addition to these 2 subjects, the training data for 1 subject were inadvertently deleted from the computer before analysis. These 3 subjects were included in the statistical analyses presented below. However, since there were no training data on which to calculate a baseline, they were eliminated from the individual prediction analyses.

Table 9 lists the results of a highly subjective assessment by the principal investigator of the quality of each individual's learning curve, broken down by age group. In subsequent studies, it will obviously be desirable to use more objective techniques to describe the individual curves (although there are no accepted techniques for doing this). However, in view of the exploratory nature of this effort, and since these subjects are not precisely representative of the FAA target population of interest, a careful subjective assessment of learning curves appeared most efficient.

In this table, "very good" indicates that the curve appeared to be of a normal shape, to plateau at a fairly early period, and to show good stability at the plateau. "Good" indicates curves showing the same characteristics, but were not quite as obvious. "Medium" indicates curves considered acceptable for testing purposes, but showed some anomaly (sporadic variability, performance degradation after plateau, etc.). Finally, "poor" indicates curves that appeared to show considerable variability, although the overall shape and ranges were considered to be acceptable, or at least marginal for individual testing decisions. As noted above, this categorization is speculative, in that no attempt was made to provide multiple assessors, or any cross-validation procedure. However, it is given here as an initial estimate to be validated by later data.

This analysis revealed that 73% of the subjects learned the task well enough to produce "very good" or "good" learning curves. An additional 15% produced "medium" quality curves, while 12% produced "poor" learning curves. Somewhat surprisingly, there were no remarkable differences in the quality of the curves as a function of age grouping. It might have been predicted that the majority of the "poor" curves might come from the oldest group. Yet, aside from a tendency to produce "good," rather than "very good" curves in this group, no such trend emerged.

**TABLE 10**

MEAN SUBJECTIVE FATIGUE SCORES BY SESSION
FOR ALCOHOL AND NON-ALCOHOL DAYS

| AGE & CONDITION | TEST SESSION 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 25 - 33 YRS | | | | | | | | | |
| NON-ALCOHOL | 5.1 | 5.4 | 6.3 | 6.6 | 7.1 | 7.9 | 8.5 | 9.7 | 10.8 |
| ALCOHOL | 4.4 | 5.4 | 6.0 | 6.5 | 7.1 | 7.7 | 8.8 | 9.5 | 11.1 |
| 40 - 47 YRS. | | | | | | | | | |
| NON-ALCOHOL | 5.0 | 7.7 | 6.3 | 7.3 | 8.4 | 9.0 | 9.3 | 11.0 | 11.9 |
| ALCOHOL | 4.8 | 5.0 | 7.2 | 7.7 | 7.7 | 8.7 | 10.0 | 11.2 | 12.2 |
| 55 - 62 YRS. | | | | | | | | | |
| NON-ALCOHOL | 4.9 | 5.9 | 6.1 | 7.1 | 7.8 | 7.9 | 9.5 | 10.6 | 11.3 |
| ALCOHOL | 5.0 | 4.8 | 7.0 | 8.1 | 8.2 | 8.7 | 9.6 | 10.3 | 10.6 |
| TOTALS | | | | | | | | | |
| NON-ALCOHOL | 5.0 | 6.3 | 6.2 | 7.0 | 7.8 | 8.3 | 9.1 | 10.4 | 11.3 |
| ALCOHOL | 4.7 | 5.1 | 6.7 | 7.4 | 7.7 | 8.4 | 9.5 | 10.3 | 11.3 |

**"Fatigue" Results**

The non-alcohol day of testing was included in the design to provide a covariate for the effects of alcohol over a full day. In other words, although this might be considered a "placebo" day for the alcohol analyses, it was also meant to reflect, in itself, the effects of a full day of "work" on the individual. In this sense, results for this non-alcohol day should reflect changes in the individual due to "fatigue," "boredom," "motivation," and any other non-alcohol factor. Results, therefore, are worthy of analysis in themselves. For convenience, the effects will be referred to as "fatigue," although it is recognized that this term is poorly defined operationally.

**Subjective Fatigue Scale**

The School of Aerospace Medicine Fatigue Scale was administered to subjects during both testing days to provide an estimate of subjective fatigue accompanying the 9 hours on task during those days. In the scoring procedure, a final "fatigue" score is obtained in which higher scores indicate greater fatigue. Table 10 presents the overall average fatigue scores obtained on non-alcohol and alcohol days by each time period, (for each age group), and Figure 5 shows the grand averaged curves on the alcohol and non-alcohol days. Inspection of these data indicates a monotonic increase in subjective fatigue as the work day progressed. There was a

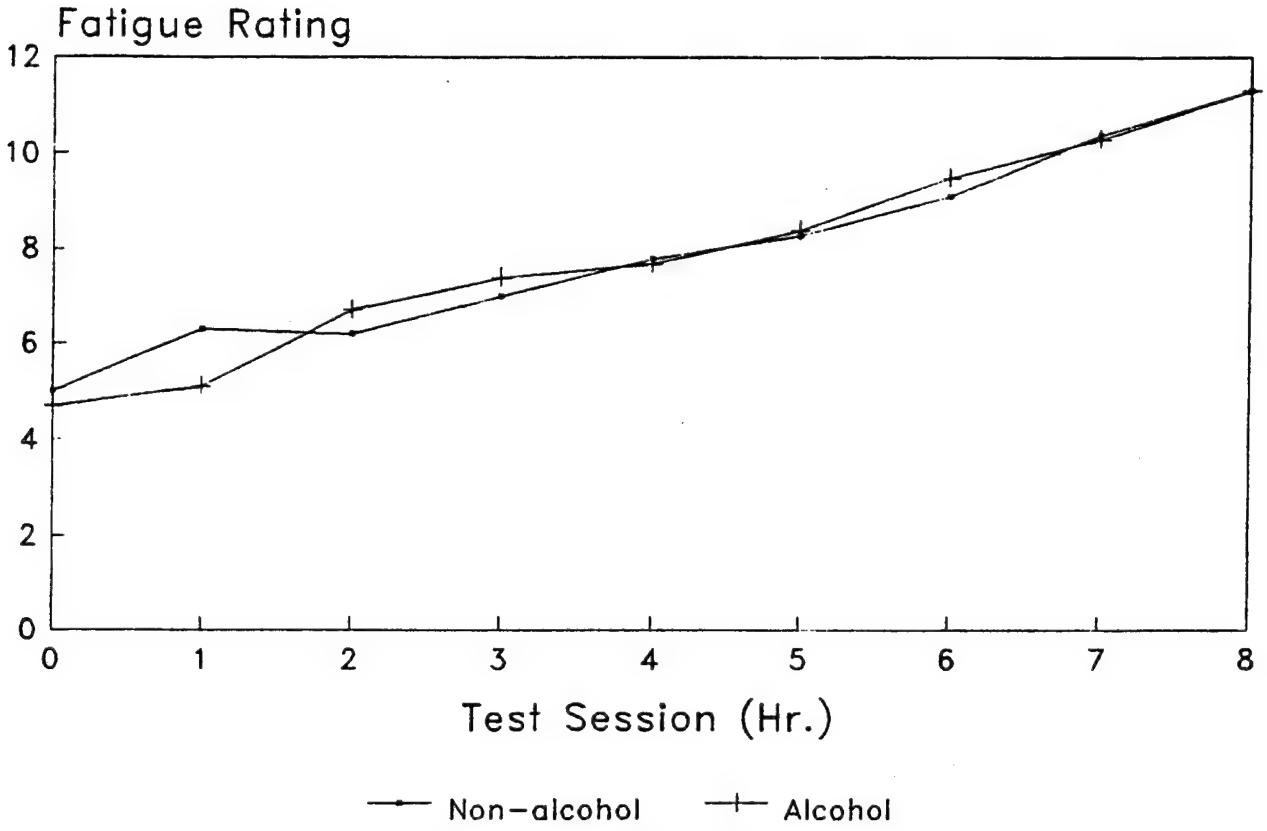



**Figure 5.** Subjective Fatigue Scale Results.

hint of a slight activation effect of the initial dose of alcohol, followed by a rebound increase in fatigue. Although this is consistent with the known short- and long-term effects of alcohol, the effects seen here are reasonably small.

These data were subjected to a 4-way ANOVA (testing sequence by age group by test session by alcohol/non-alcohol). This analysis revealed a highly significant main effect of test session ($p<.0001$), and an interaction effect between testing sequence and alcohol/non-alcohol ($p<.007$). Thus, although the consistent increase in self-reported fatigue during the course of each day was highly reliable, there were no significant differences as a function of alcohol or age. The apparent short- and longer-term effects of alcohol on fatigue were not demonstrable statistically.

Post-hoc analyses of the interaction effect indicated that when subjects had their alcohol day first, they showed more fatigue (virtually all day) on their **placebo** day (their second test day). A similar, but less strong effect was seen when subjects had their placebo day first. In that case, they showed slightly more fatigue on their **alcohol** day. In other words, subjects reported more fatigue on the **second** day of testing, regardless whether it was an alcohol or non-alcohol day. This could have been due to a motivational effect, or simply an effect of becoming more familiar with (and bored with) the reporting forms. In any case, it raises the possibility that there were basic attitudinal and affective changes in the subjects from the first to second day of testing that might be independent of the experimental manipulations.

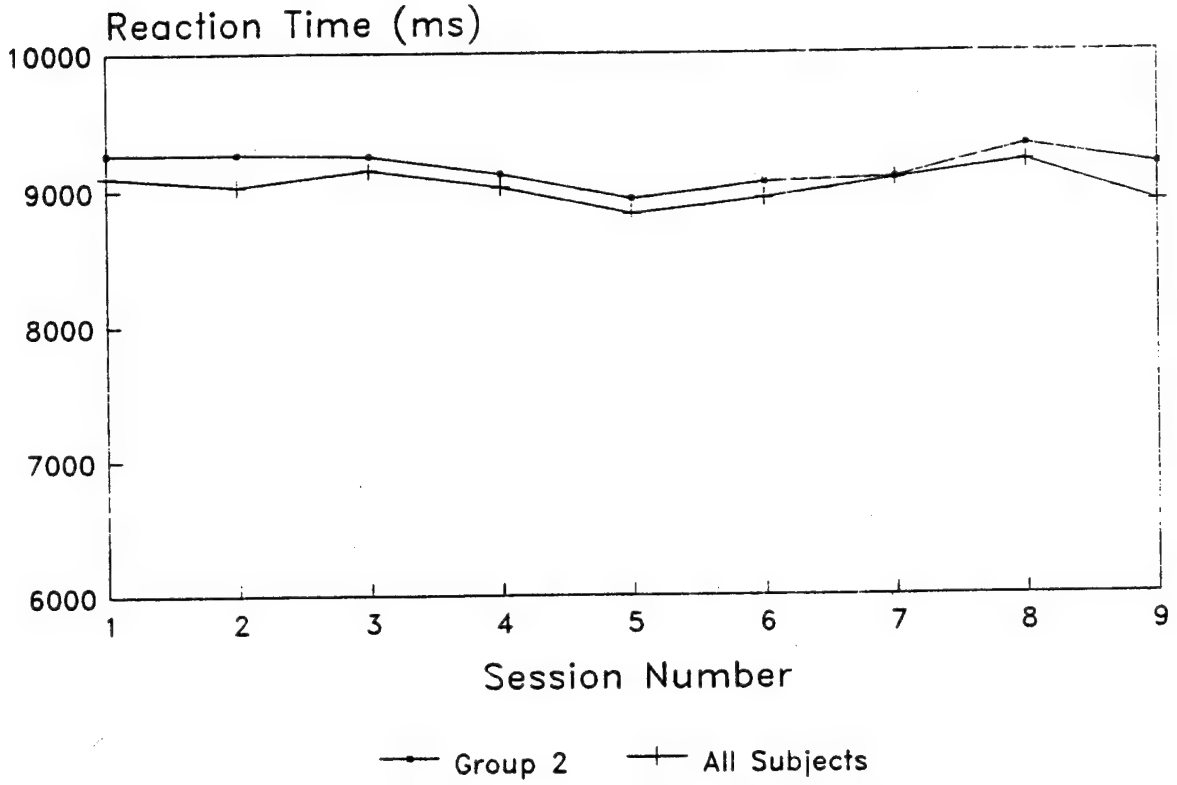



**Figure 6.** Representative FAA-RTP Performance "Fatigue" Curves for All Subjects, and Just for Subjects Experiencing the Non-Alcohol Day First (Group 2).

**TABLE 11**

SIGNIFICANT ANOVA RESULTS FOR ALL FAA-RTP VARIABLES ON THE NON-ALCOHOL ("FATIGUE") DAY

| VARIABLE | EFFECT | p-LEVEL |
|---|---|---|
| T1RT | AGE | .0010 |
| | AGE x SEQUENCE | .0009 |
| T1TP | AGE | .0009 |
| T1TRRT | AGE x SEQUENCE | .0033 |
| T2RT | AGE | .0009 |
| T2TP | AGE | .0017 |

**FAA-RTP Measures of Fatigue**

The subjects' FAA-RTP performance on the non-alcohol day was plotted as a function of test session. The 15 curves obtained showed consistently "flat" functions, suggesting that there was no fatigue effect on FAA-RTP variables demonstrable on groups in this study. Representative examples of these curves, for both the total subject sample (A), and also for those who had the non-alcohol day first (B) are shown in Figure 6. These non-alcohol day data were analyzed by independent 3-way ANOVAs for each of the 15 FAA-RTP variables. In this and all subsequent univariate analyses presented in this report, significance values have been adjusted by using the Bonferroni inequality (Wilcox, 1987) to account for the multiple analyses performed in the data. The main factors were 1) test sequence (alcohol day first or fatigue day first), 2) age group, and 3) session (test session on a given day). The significant results are summarized in Table 11. Four of the variables showed a main effect of age at the Bonferroni-protected level $<.05$ (requiring $p$ values less than .0033), and 2 showed an interaction between age and test sequence. In all cases, post-hoc (Newman-Keuls) tests revealed that the age effect was because the older group did worse than either the younger or middle-age group. The interactions were generally because the younger group which received alcohol first, did significantly better than most other groups. No variables showed session effects.

These results establish the fact that the older group performed significantly worse than either of the other age groups, independent of any fatigue effect. The interaction with test sequence may suggest that there is still "learning" going on in the younger group, even after an apparent "plateau" has been reached. This is possible since the subjects who had the placebo day first had 9 more "training" sessions (under undegraded conditions) than those who had the alcohol day first. However, this must be treated as a tentative finding, since none of the other age groups showed the effect.

The failure to find significant effects of sessions on FAA-RTP suggests that the test did not statistically demonstrate any fatigue decrement in performance over a 9-hour period. However, the finding that some "learning" may have continued in the younger group over the course of the fatigue sessions might have confounded this result. This continued improvement could have been enough to counter any true fatigue main effect, even though continued improvement apparently did not occur in the other 2 age groups. Obviously, individuals were subjectively experiencing fatigue, as demonstrated by their subjective ratings. In any case, for whatever reason, the levels of fatigue experienced apparently were not sufficient to produce group decrements on FAA-RTP.

**Statistical Analyses of FAA-RTP Alcohol Results**

***Overview***

The basic question of this study was whether the FAA-RTP test is able to differentiate among subjects who have no alcohol in their system, and those who have various levels of alcohol. The goal is to ultimately develop a "dose-response" curve for the FAA-RTP test with respect to BrAC. To answer this question, 2 interrelated approaches must be taken. First, it is desirable to demonstrate whether FAA-RTP variables differentiate **statistically** among the experimental conditions. To do this, multivariate analyses of variance (MANOVAs) were first performed to determine whether the pre-selected set of FAA-RTP measures reliably differentiated between experimental conditions. Where justified by these results, univariate and appropriate post-hoc analyses were then inspected to probe the nature of such differences.

It is important to remember that, while this is not an exploratory study in a technical sense, it is devoted to **developing** scoring procedures for a new implementation of the FAA-RTP procedure. Therefore, some degree of exploration of the data is justified, rather than rigid statistical adherence to predictions. Nevertheless, it is also obviously important to protect against capitalization on chance, especially where a large number of measures are available. MANOVA procedures (Wilcox, 1987) and (to provide additional protection) the use of the Bonferroni correction (Morrison, 1976) to protect the pre-established significance probability of .05,

**TABLE 12**

FAA-RTP MEASURES USED IN MANOVAs

| |
|---|
| Vector Reaction Time |
| Vector Transition Reaction Time |
| Matrix Task Reaction Time |
| Matrix Task Transition Reaction Time |
| Angles Task Reaction Time |
| Vector Percent Correct |
| Vector Transition Percent Correct |
| Matrix Task Percent Correct |
| Matrix Task Transition Percent Correct |
| Angles Task Percent Correct |

operate to reduce such chance results. Ultimately, however, no statistical procedure alone is comparable to a well designed cross-validation study. In the interim, the statistical analyses presented below provide the firmest foundations available for future hypotheses and study.

While the above analyses are necessary to demonstrate the statistical reliability of any effects found, they do not address the power of FAA-RTP to discriminate among **individuals** with respect to alcohol. To do this, an individual scoring "algorithm" must be created for this new test, and the power (sensitivity and specificity) of the algorithm must be tested at various levels of BrAC. The present study was not designed primarily for this purpose. However, some **preliminary** estimate of the test's efficiency and power can be obtained. It can not be overemphasized, however, that these preliminary analyses are just that —preliminary— they can provide important clues for further hypotheses and study, but they should not be over-interpreted or used to establish final procedures or conclusions.

***Raw Data Analyses***

Considering the numerous FAA-RTP measures, test sessions, and sequences of independent variables, it was first necessary to carry out multivariate analyses to determine if there were overall general effects. For these multivariate analyses, 72 subjects were analyzed. Further, to reduce the data set to a manageable size, only the ascending limb of the alcohol curve (i.e., the first 5 test sessions), and only 10 FAA-RTP variables (see Table 12) were analyzed with multivariate statistics. The analyses in this section were performed with SYSTAT for Windows V5.2 (1992).

To help assess the presence of compound symmetry, the Huynh-Feldt statistic (see SYSTAT, 1992) was used. This is used to adjust the probability for the classical univariate tests when compound symmetry fails. As a rule of thumb, if the Huynh-Feldt $\underline{p}$ values are substantially different from the standard univariate probability values, then one should be suspicious that compound symmetry has failed. Inspection of the data revealed that the Huynh-Feldt values were generally close to the univariate $\underline{p}$ values (within .01 in all cases). Thus, the more sensitive univariate measures could be used. However, for the present purpose, the more conservative approach of using only the multivariate output was chosen.

Three significance statistics can be computed for each multivariate analysis, each addressing different concerns about underlying assumptions. In the present analyses, all 3 were computed. Wilks' Lambda (likelihood ratio criterion) varies between 0 and 1. The F statistic for Wilks' Lambda is Rao's approximate F statistic corresponding to the likelihood ratio criterion (see SYSTAT, 1992). Pillai's

**TABLE 13**

SIGNIFICANCE LEVELS OF MULTIVARIATE ANALYSIS OF VARIANCE

| TEST OF | p |
|---|---|
| FAA-RTP MEASURES | 0.040 |
| SESSION | 0.001 |
| ALC/PLACEBO | * |
| SEQUENCE | * |
| MEASURES*SESSION | 0.497 |
| MEASURES*ALC/PLACEBO | 0.035 |
| MEASURES*SEQUENCE | 0.001 |
| SESSION*ALC/PLACEBO | 0.001 |
| SESSION*SEQUENCE | 0.282 |
| ALC/PLACEBO*SEQUENCE | * |
| MEASURES*SESSION*ALC/PLACEBO | 0.524 |
| MEASURES*SESSION*SEQUENCE | 0.004 |
| MEASURES*ALC/PLACEBO*SEQUENCE | 0.001 |
| SESSION*ALC/PLACEBO*SEQUENCE | 0.290 |
| MEASURES*SESSION*ALC/PLACEBO*SEQUENCE | 0.004 |

* Not computed for variables with 1 df

trace and its F approximation were described by Pillai (1960). The Hotelling-Lawley trace and its F approximation are described by Morrison (1976). The full tables for each of these statistics are presented in NTI, Inc. (1993). In fact, the F values and p levels for the 3 tests were always essentially the same for every table entry. Therefore, only the common p values for these tests are shown in Table 13 below. Note that there is no table entry for the main effect of sequence of testing (alcohol day first or second) or for alcohol vs. placebo day. These procedures do not compute results for variables with only 1 degree of freedom.

Inspection of this Table reveals that there is a main effect due to MEASURE ($p < .04$)). This is not surprising since reaction time and percent correct values were included in the mix. More importantly, there is a main effect due to SESSION ($p < .001$). This confirms, of course, that there were changes in scores as BAC levels increased. The significant interaction between MEASURES and ALC/PLACEBO condition ($p < .035$) was a hypothesized result, since the various measures were expected to differ in their ability to differentiate alcohol from placebo conditions. A most important result is the clear interaction between SESSION and ALC/PLACEBO ($p < .001$). Generically, this result appears because the test scores changed differently over the alcohol day than they did over the placebo day. Finally, there are a number of significant higher order interactions with the SEQUENCE in which alcohol and placebo were given. Although it is extremely difficult to interpret higher order interactions in a MANOVA, this suggests that order had a strong effect on many variables. Although these did not appear to interfere with effects of alcohol and

## TABLE 14

### SIGNIFICANT ANOVA RESULTS — FAA-RTP RAW SCORES

| FAA-RTP VARIABLE | ANOVA MAIN AND INTERACTION EFFECTS | | | |
|---|---|---|---|---|
| | AGE | SESSION | SEQUENCE X ALC INTERACT. | SESSION X ALC INTERACT. |
| T1RT | .0006 | | .0001 | .0007 |
| T1TP | | | .0001 | |
| T1TRRT | .0020 | | .0001 | .0010 |
| T1TRTP | | | .0001 | |
| T2RT | .0006 | .0012 | .0001 | |
| T2TP | .0016 | | .0001 | |
| T2TRRT | .0029 | | .0001 | |
| T2TRTP | | | .0001 | |
| ATTNRT | | .0001 | | .0022 |
| ATTNTP | | .0001 | | .0030 |

sessions, they point up the fact that in subsequent studies, additional care must be exercised to assure that sequence of testing effects, learning, and other factors of peripheral concern are better controlled.

In view of the significant MANOVA results, a series of 4-way univariate analyses were performed on each variable separately. The main factors of these analyses were 1) test SEQUENCE, or whether the alcohol or non-alcohol day was first, 2) AGE, 3) SESSIONS, and 4) alcohol or non-alcohol day — ALC/PLACEBO.

The first set of data analyzed consisted of the raw scores produced by the subjects during the test days. For these analyses, all subjects were included, and no attempt was made to remove the few subjects who had not achieved a stable baseline or who, for other reasons, might have been questionable. Significant results of these first analyses are presented in Table 14, and if an effect is not shown, it can be assumed to be non-significant. To simplify the presentation, only variables of greatest interest are presented here and in subsequent tables. The full ANOVA Tables are presented in NTI, Inc. (1993).

These results suggest that 5 of the FAA-RTP variables were affected by AGE. Newman-Keuls post-hoc tests on these variables yielded a universal effect in which the older group did significantly worse than either of the younger groups. However, the general lack of significant interactions with this factor suggests that the age-performance difference did not affect either the test's sensitivity to alcohol or to fatigue.

No significant main effects on any variable were found for SEQUENCE of testing, and no interaction effects were found for SEQUENCE by SESSION, AGE by SESSION, AGE by ALC/PLACEBO, or any higher order interactions. The significance of the interaction between testing SEQUENCE and ALC/PLACEBO could also be meaningful. Post-hoc tests revealed no consistent pattern,

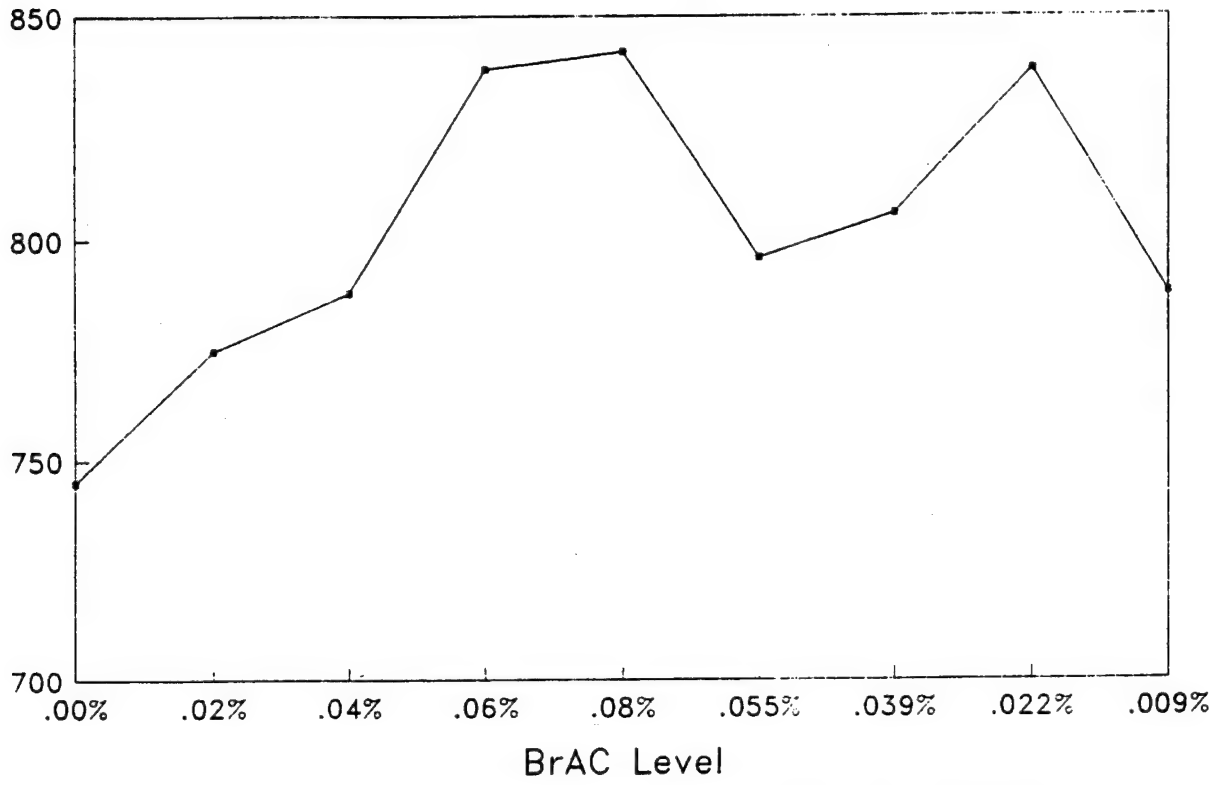



**Figure 7.** The Angles Reaction Time Measure of the FAA-RTP Test.

which would suggest that there was a true sequence effect, or that any such minor effect was consistently related to alcohol consumption. For instance, for several variables the group receiving alcohol first did worse on their alcohol day than the non-alcohol-first group did on their alcohol day. For a few variables, this pattern was reversed.

Lacking any meaningful pattern, the effect must be attributed either to practice, or to a motivational change from day 1 to day 2 of testing. In any case, these data point up the desirability of removing this source of variance from subsequent analyses.

The effects of main concern in the present case are the SESSION effect and, especially, the interaction of SESSION with the ALC/PLACEBO variable. The former establishes that there was variation over sessions (even when alcohol days were averaged with non-alcohol days), and the latter can isolate the actual effect of the alcohol. Three of the 15 variables were significant for the main effect of sessions. These three were the matrix task reaction time (T2RT), the angles task reaction time (ATTNRT) and thruput (ATTNTP). Two variables (both involving the vector task) were significant for the interaction. Thus, each of the 3 FAA-RTP tasks showed some statistical differences between alcohol and non-alcohol conditions.

For the main effect of sessions, the strongest effect appeared in the angles task, with both reaction time (ATTNRT) and thruput (ATTNTP) showing significance beyond .0001. Post-hoc tests revealed that thruput in the zero alcohol condition was significantly different from .04% BrAC ($p < .01$), .06% BrAC (.02), .08% BrAC (.0007), and also all BrAC conditions on the descending limb of the alcohol curve. In addition, the session by alcohol/placebo interaction was also significant for these 2 variables. Thse effects are shown graphically in Figure 7.

In summary, analyses of these raw data indicate that some variables of the FAA-RTP test vary as a function of BrAC, even without considering the subjects' individual baselines. Specifically, at least 1 variable discriminated between zero and .04% BrAC at the .01 level of significance. Several other

variables also discriminated between alcohol and non-alcohol conditions at various concentrations. All 3 tasks in the FAA-RTP test showed some statistical sensitivity to alcohol.

***Deviation Score Analyses***

When the FAA-RTP test is used in industrial settings, an individual's "baseline" and normal range of variability are calculated from the 15 to 30 test sessions after the person has reached plateau. The difference between the baseline and the score on any given day would then be divided by the baseline variability. The shift to within-subjects analysis tends to remove bias due to individual differences.

To better test the FAA-RTP test under conditions closer to those which would be employed in an actual implementation, a modified form of the above procedure was carried out on the present data. The modification consisted of using a smaller number of sessions to estimate the mean and standard deviation than would be used in actual practice. This was necessitated by the fact that relatively few post-plateau sessions were available, due to the experimental design. Therefore, each individual training curve was inspected, and all of the relatively stable sessions after plateau were designated as the "baseline" sessions. This resulted in as few as 4 and as many as 13 sessions being used to calculate the baseline (mean = 7). The net effect of this modification was to make the baseline estimates less reliable and stable than they would be in actual practice, thus introducing the possibility that more test variability would be observed in this study than would be expected in actual field implementations.

**TABLE 15**

SIGNIFICANT ANOVA RESULTS — FAA-RTP DEVIATION SCORES

| FAA-RTP VARIABLE | ANOVA MAIN AND INTERACTION EFFECTS | | |
|---|---|---|---|
| | SESSION (3) | SEQUENCE X ALC INTERACT. | SESSION X ALC INTERACT. |
| T1RT | | .0001 | .0001 |
| T1TP | | .0001 | .0001 |
| T1TRRT | | .0001 | .0001 |
| T1TRTP | | .0001 | .0001 |
| T2RT | .0017 | .0001 | |
| T2TP | | .0001 | |
| T2TRRT | .0001 | | |
| T2TRTP | | .0001 | |
| ATTNRT | .0001 | .0022 | |
| ATTNTP | .0001 | .0031 | |

To summarize the data treatment, each subject's mean (Mbaseline) and standard deviation (SDbaseline) over the designated number of baseline sessions was used as his "baseline." The subject's raw score on each variable for each test session (X) was then subtracted from the mean baseline score (Mbaseline) and this result was divided by the baseline standard deviation (SDbaseline). The resultant score was used as the subject's score for each testing session (Xz):

$$Xz = \frac{M_{baseline} - X}{SD_{baseline}}$$

The Xz values then served as the raw data for the 4-way ANOVAs summarized in Table 15, and the ANOVA Tables are presented in NTI, Inc. (1993).

The first notable effect of this transformation was to remove virtually all of the "age" effects seen in the raw data analyses. Of course, this could be expected, since the subjects were now being evaluated against their own baseline. Since age did not interact with the other study variables in the previous analyses, it is logical that the effects would disappear with the present approach.

On the other hand, the effect due to the interaction between sequence of testing and alcohol/non-alcohol day did not disappear. These analyses help clarify similar results seen above in the raw data analyses. Whereas there was no clear pattern in those data, post-hoc analyses of the deviation scores suggested that there was a genuine improvement in performance from the first day of testing to the second, independent of alcohol condition. Again, whether this was due to a practice effect, or to some motivational effect, is problematical. However, these data again point up the necessity to assure that very stable baselines are achieved in future studies before "test" data are collected.

The main effects of session continued to show several FAA-RTP variables to be significant. These were all reaction time variables. Again, the angles task showed the most reliable statistical results. Post-hoc tests revealed that this effect for the thruput measure (as well as others) was concentrated on the difference between the zero condition and the .08% BrAC condition ($p < .0007$) on the **ascending** limb of the BrAC curve. On the descending limb, there were significant differences between the zero alcohol condition and **all** other alcohol conditions ($p$ values ranging from .0001 to .03). The actual values for this measure are illustrated in Figure 8. They reveal an interesting pattern of scores that show a monotonic decrease in performance with increasing BrAC for the **ascending limb of the curve**. However, when BrAC is decreasing, the pattern is more complex. There is an initial improvement (perhaps related to the fact that subjects ate some food after their .08% BrAC testing session). However, performance then **declines**, even though BrAC is still **decreasing**. In fact, performance is most degraded at the 8-hour point, when average BrAC is only at .022%. This may reflect the interaction of alcohol with fatigue, or may suggest that the descending limb of the alcohol curve has more variable effects on performance than the ascending limb.

The significance of the interactions between sessions and alcohol was stronger for the deviation scores than it had been for the raw scores. Post-hoc tests revealed that this was primarily due to differences between the zero condition and the .08% BrAC condition on the ascending limb of the curve, as well as some differences between zero and points on the descending limb. Based on this, it is reasonable to hypothesize that some combination of these measures could at least identify subjects at .08% BrAC on the ascending limb, and at lower levels on the descending limb.

Other analyses of these data were carried out to determine whether different data treatments could reveal more about the FAA-RTP test. Specifically, log transformations of the reaction time data were performed, and all statistical analyses were re-run to determine whether the increased sensitivity to the direction of performance change would affect results. A description of these analyses is presented in NTI, Inc. (1993). Essentially, both treatments would have resulted in identical decisions in those factors of major interest.

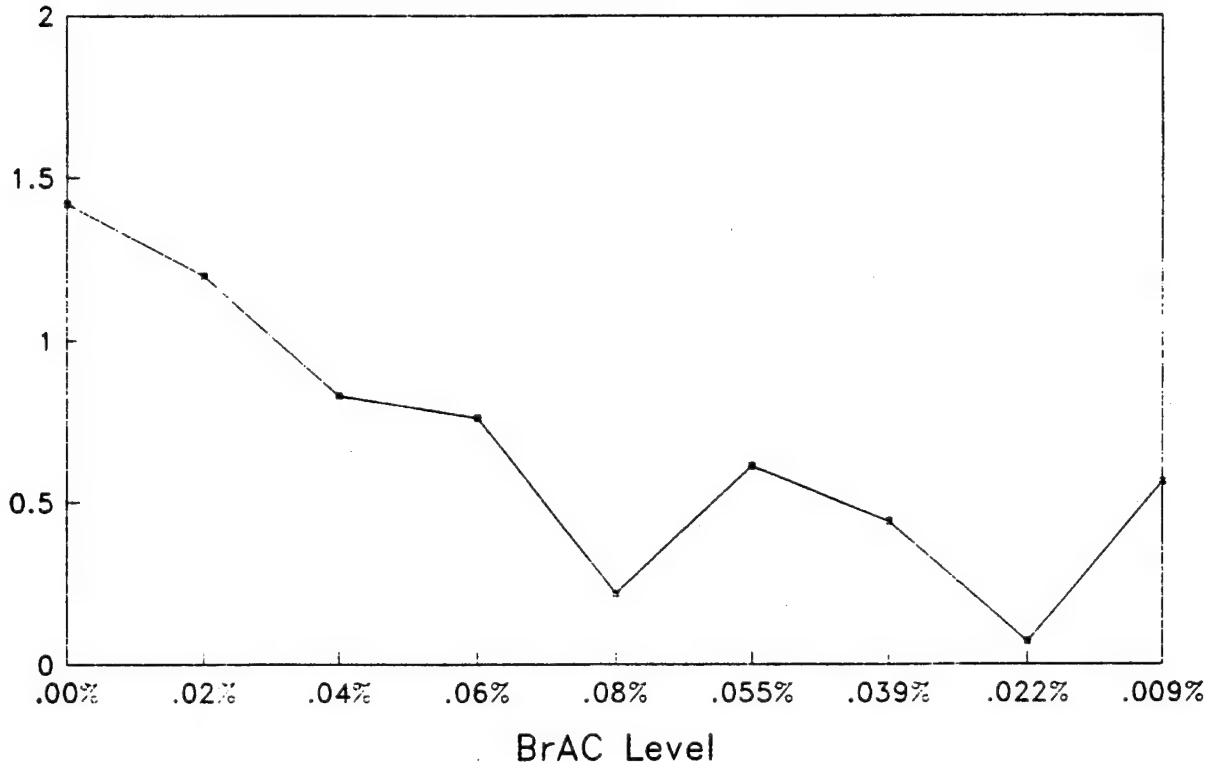



**Figure 8.** FAA-RTP Thruput Deviation Scores as a Function of Alcohol Level.

In addition to the above, a factor analysis was performed on the test variables, and details are reported in NTI, Inc. (1993). As expected in the situation where percent correct is consistently very high, reaction time and thruput measures were highly correlated. Four factors accounted for 68% of the variance. Two of these were relatively independent factors for reaction time measures and measures on the attention task, and two involved combinations of percent correct measures.

**Analyses of Individuals' Scores on the FAA-RTP Test**

Since the ultimate purpose of the FAA-RTP procedure is to test individuals, there is, naturally, great interest in the ability of the test to detect given alcohol levels in the individual. A test may show statistical sensitivity, and yet, not meet criteria for successful individual prediction. Eventually, the FAA-RTP test must demonstrate that it can be used to detect performance impairment due to alcohol (as well as other causes) with a high degree of sensitivity and specificity. Therefore, the present data were inspected from this viewpoint, and the results are summarized below.

However, several serious cautions must be pointed out before this analysis is presented. The first problem deals with the nature of the present experiment. This study was designed to isolate and define the **statistical** sensitivity of the FAA-RTP test to alcohol. The procedures employed are, therefore, based on an experimental design, not on the way FAA-RTP would be used in the "real world." Most notably, the baselines used in the present study consisted, generally, of 4 to 10 sessions, which are considered sufficiently stable to demonstrate statistical effects. In actual practice, where individual discrimination is desired, between 15 and 30 sessions would be used to calculate these baselines. Therefore, individual analyses in the present study would be expected to show more variability than would be seen in actual implementations. Other

factors, such as subject selection, and experimental demand effects, which can be controlled in statistical analyses of experimental data, contribute to uncontrolled variation in individual analyses.

A second, related problem stems from the fact that the analysis below is absolutely data specific. No cross-validation was planned or carried out on these data. There is no doubt that such a cross-validation would show some shrinkage of the results seen here. Thus, the results below must be viewed as totally suggestive, and not definitive in any sense.

A caution should also be pointed out concerning the use of the term "false positive" below. With respect to a readiness-for-duty test, a false positive refers to an individual who is not degraded, but who "fails" the test. If the person fails the test and is actually degraded, from any of a **variety** of causes, the result is a true positive. In the following discussion, however, the term is used specifically with respect to the breath **alcohol** burden of the person. An individual is considered a false positive if he fails the test, even though he has a BrAC of zero. In fact, such an individual may actually have been degraded from some other cause. In such a case, he would not be a false positive, but a true positive in the broad sense of the term. For these reasons, the term "false positive" (below) should be understood to mean "false positive **with respect to alcohol burden**." No generalizations concerning the probable incidence of general false positives in an actual test implementation can therefore be drawn from these data.

Finally, a most serious problem rests in the very fact that a performance test is being used to detect blood alcohol levels. The precise correlation between blood alcohol level and performance decrement is unknown, but is certainly not perfect. Although government agencies have mandated that certain blood alcohol levels are to be prohibited, it is universally accepted that one can not **assume** that there will be a performance decrement at any particular time at lower BrAC levels. Therefore, since the FAA-RTP test is designed to measure **performance**, there is no reason to assume, **a priori**, that it will be perfectly correlated with blood alcohol levels.

In spite of the above cautions, it is reasonable to probe the individual sensitivity of the test to alcohol, even when conditions are not realistic or ideal. Such individual sensitivity should be based on the ability of the test to "detect" each of the experimentally determined BAC levels. Data dealing with this question are presented below.

Given the fact that baselines in the present study will be intrinsically less stable than those which will be collected in an actual implementation of the FAA-RTP test, it is reasonable that any subjects who did not show an adequately stable baseline for individual analysis should be eliminated from the analysis. Obviously, subjects who produced learning curves which were considered "poor" in the original classification (see Table 9) were automatically eliminated. Similarly, those who had fewer than 16 training sessions were eliminated because there was not enough data to assume stability. This procedure resulted in the elimination of 19 subjects from the individual analyses. In addition, 14 other subjects showed more variability during the baseline sessions than would be allowed during an actual implementation. The remaining 41 subjects, while probably not quite as stable as one would demand in practice, were considered usable for individual analyses.

The procedure for determining "pass-fail" criteria for any new implementation of the FAA-RTP procedure involves inspection of the statistical significances found in sensitivity studies to determine: 1) which variables hold most promise for providing individual detection criteria, and 2) whether there are **one or more patterns** of variables that appear to differentiate various alcohol levels. For this study, this process was carried out by visual inspection (although, ultimately, more sophisticated statistical and non-linear mathematical analyses must be used).

From this inspection, candidate "scoring algorithms" were created. These algorithms were based on hypothesized effects of alcohol on various measures available from FAA-RTP. They were applied to the data for each subject at each alcohol level, and a "true positive" rate was calculated (defined as an individual at a given alcohol level who

"failed" the test). A "false positive" was defined as an individual at zero BAC (on either the alcohol or non-alcohol day) who failed the test. "True negatives" and "false negatives" were similarly defined for individuals who passed the test and 1) who had 0 BAC, or 2) who had a given level of alcohol, respectively.

In fact, a simple approach, similar to that used frequently in clinical medicine, proved to be essentially as effective as more complex scoring algorithms. The major criterion of "passing" and "failing" was simply **whether the individual fell below the cut-score on any 2 or more of the 15 variables**. When this criterion was applied to the data, using a cut-score of 2.00 SD, sensitivity and specificity rates shown in Table 16 were obtained. The 2.00 SD cut-score represents a "conservative" scoring approach (defined as that which would detect the fewest number of false positives, at the risk of detecting fewer true positives). In past NovaScan applications, such a conservative cut-score has correlated very well with actual job performance decrement (O'Donnell, 1993a).

Inspection of Table 16 suggests that about 76 to 78% of individuals with BrAC between .02% and .04% would be detected with this scoring approach. At .06% BrAC, this figure rises to 88%, and reaches 97% at .08% BrAC. The 1 individual who was not "failed" at .08% BrAC did, in fact, fail on the matrix test, but this did not qualify as a "fail" under this scoring criterion.

The above findings must also be tempered somewhat by the relatively large number of false positives (30%). In actual RTP implementations, a test is always given at least twice before an individual is "failed." Analysis of the first 2 tests given on the non-alcohol day revealed that 10 of the 41 individuals failed both tests (24%).

**TABLE 16**

SENSITIVITY AND SPECIFICITY* VALUES FOR THE FAA-RTP USING A CUT-SCORE OF -2.00 SD ON TWO VARIABLES

CUT-SCORE: -2.0 STANDARD DEVIATIONS BELOW BASELINE
SCORING CRITERIA: FAIL ON ANY TWO FAA-RTP VARIABLES

| | ALCOHOL LEVEL | | | | |
|---|---|---|---|---|---|
| | 0 | .02 | .04 | .06 | .08 |
| FAA-RTP "PASS" | 58 | 9 | 10 | 5 | 1 |
| FAA-RTP "FAIL" | 25 | 32 | 31 | 35 | 38 |
| SENSITIVITY | | 78% | 76% | 88% | 97% |

SENSITIVITY (Based on initial test on each day) = 70%

* Sensitivity is defined in the traditional way as ratio of the number of true positives to the total number of positives. Specificity is similarly defined as the ratio of the number of true negatives to the total number of negatives.

If one adopts a less conservative approach with respect to the cut-score (i.e., is willing to accept more false positives), it would be expected that the sensitivity rates would rise and specificity would fall. To explore what would happen under these conditions, the cut-score was changed to 1.5 standard deviations below the subject's baseline, and **the subject failed if any 2 variables fell below that cut-score**. In addition, 6 specific variables that appeared to be good discriminators were added to the algorithm. If the subject failed any of the following variables (at a cut-score of 1.0 standard deviations), the test was considered a "failure:"

1. Vector task reaction time
2. Vector task transitions reaction time
3. Matrix task transitions reaction time
4. Matrix task transitions percent correct
5. Angles task reaction time
6. Angles task thruput

Using these "liberal" criteria, the data shown in Table 17 were obtained. These results indicate that 100% of subjects at .04% BrAC can be identified with the FAA-RTP procedure. In fact, the test detected 92% of the subjects at .02% BrAC.

Table 17 reveals sensitivity values between 90 and 100% at all BrAC levels tested. As noted previously, since the correlation between alcohol level and performance is not perfect in the lower ranges, it would be expected that there would be some variability in pick-up rates, and in fact a "reversal" is seen between .04% and .06% BrAC, with the higher BrAC producing 4 false negatives. There could be several practical reasons for such a reversal (e.g., a subject who failed marginally at the .04% level might have realized this and "mobilized" resources on the next test enough to marginally pass). However, such speculation is probably not warranted at this time. This is the first attempt to arrive at a scoring algorithm for the FAA-RTP test, and these results must be considered data-specific until cross-validation can be carried out. For these reasons, the above reversal again emphasizes the fact that the present results should not be interpreted too literally. They only indicate that FAA-RTP test performance appears to degrade in the majority of individuals at BrAC levels that are considered to be very low.

This sensitivity is purchased at the cost of considerably reduced specificity. With the more liberal scoring algorithm, about 44% of the subjects with no alcohol in their system were incorrectly classified on the first tests. This figure was reduced slightly (to 41%) if 2 tests were considered. Obviously, if these were actual false positives, these rates would not be acceptable in an actual implementation of the test. However, again, it is likely that some unknown portion of those who were classified as false positives **with respect to alcohol** were actually degraded from some other cause. These individuals would therefore be considered true positives.

In summary, these individual analyses confirm that the FAA-RTP test developed for the FAA is **capable** of detecting 100% of valid test subjects at BrAC levels of .04%, if a reasonably high false positive rate is acceptable. Given the fact that the present experimental design probably operated to introduce more variability into these data than would be seen in actual field operations, it can be concluded that the test can be sensitive to low levels of alcohol on an individual basis.

**Questions Concerning the Timing of the Test**

In the initial design of this project, questions were raised involving the amount of time it would take to administer the FAA-RTP test. This is a critical practical question that severely limits the utility of any testing procedure. A target limit of 10 minutes was established by the FAA at the inception of this effort. Since the FAA-RTP test is a new device, there were no data on the testing times required, or even on the minimum number of stimuli required to achieve the desired level of discrimination among subjects. For these reasons, calculations were made on the time taken by subjects to complete the test, and an attempt was made to determine whether the number of stimuli used in this experiment could have been reduced. These results are reported in this section.

**TABLE 17**

SENSITIVITY AND SPECIFICITY VALUES FOR THE FAA-RTP
USING A CUT-SCORE OF -1.5 SD ON TWO VARIABLES
PLUS -1.0 SD ON SIX SPECIFIC VARIABLES

CUT-SCORE: -1.5 STANDARD DEVIATIONS BELOW BASELINE
SCORING CRITERIA: FAIL ON ANY TWO FAA-RTP VARIABLES
OR FAIL ON 6 SPECIFIC VARIABLES

| | ALCOHOL LEVEL | | | | |
|---|---|---|---|---|---|
| | 0 | .02 | .04 | .06 | .08 |
| FAA-RTP "PASS" | 46 | 3 | 0 | 4 | 0 |
| FAA-RTP "FAIL" | 36 | 38 | 41 | 36 | 39 |
| SENSITIVITY(%) | | 92 | 100 | 90 | 100 |

SENSITIVITY (Based on initial test on each day) = 56%

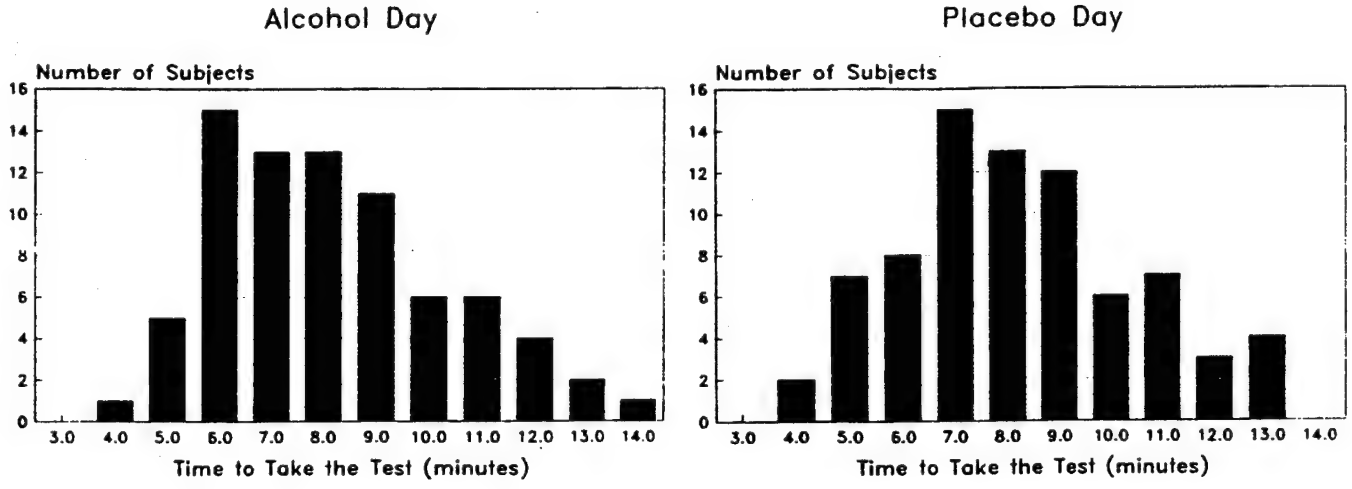



**Figure 9.** Total Time to Take the FAA-RTP Test on Both the Alcohol and Non-Alcohol ("Placebo") Day.

**Time Required for a Single Test**

Since all subjects' testing times were recorded from the computer clock, it was a relatively simple matter to determine the total time taken by the subject to complete a single FAA-RTP test. The results presented in Figure 9, therefore, refer to the total elapsed time from the subject starting the actual test (excluding warm-up) to the time the computer saved the data. The warm-up time was not included in these calculations because the number of warm-up trials which will ultimately be provided, has not been determined. In the present experiment, 24 warm-up trials were provided. Thus, one could add at least 13% to the times indicated below (24/80) to account for the warm-up time. However, in actual practice, as few as half of these warm-up trials will probably prove adequate, so the total time would be reduced proportionally.

Clearly, there was a shift toward faster completion of the test on the alcohol day, with modal values of 7 minutes on the placebo day and 6 minutes on the alcohol day. The median values and interquartile range of these 2 distributions are shown in Table 18.

Although interesting, the differences between the alcohol and placebo day are small, and probably not significant from an operational point of view. Seventy-five percent of all subjects were able to complete the test in under 9 minutes, and fewer than 19% of these unselected subjects took longer than 10 minutes to complete it. This suggests that a more selected population might show an even lower median test time, and that there would be a much lower percentage who might exceed 10 minutes.

The above data were collected on trained subjects who were performing near their operational limit of speed. It should be noted that when subjects first start training on the FAA-RTP test, the testing times were much longer. It was not unusual for a subject to take 20 minutes or more to complete 1 test during the first day of training. For a few of the subjects, this pattern continued for 10 or more training sessions. The majority of subjects, however, rapidly decreased their per/test time during training, and achieved their final speed by the 10th to 16th training session. Although no objective records were analyzed during training, it is probably safe to assume that, on average, training trials will be 9 to 11 minutes long, with a wide variation both within and between subjects.

When the above timing data were inspected as a function of age group, a slight tendency was observed for the older group to take longer to complete the test. However, this was a very mild trend on both the alcohol and placebo days, and probably would not affect the aggregate logistics of administering the test to large groups.

**TABLE 18**

MEDIAN AND INTERQUARTILE RANGE TIMES TO TAKE THE FAA-RTP TEST

| MEASURE | PLACEBO DAY | ALCOHOL DAY |
|---|---|---|
| MEDIAN | 7.50 min. | 7.36 min. |
| 1st QUARTILE | 6.15 min. | 6.90 min. |
| 3rd QUARTILE | 8.10 min. | 8.90 min. |
| RANGE | 1.95 min. | 2.00 min |

**Sensitivity with Reduced Numbers of Stimuli**

The number of trials presented to a subject, of course, determines the length of the test for that subject. Since it is desirable to use the smallest number of trials that would give statistically reliable and practical results, an analysis was carried out to determine whether a test using 60 trials-per-task would have produced the same results as the 80 trial-per-task used in the present experiment.

To accomplish this, new baselines were calculated for each subject, based on the first 60 stimuli for the sessions previously identified as "baselines." It should be noted that this procedure essentially destroyed the counter-balancing of stimuli described previously for the 80 stimulus test. It is, therefore, possible that the results presented below could be confounded, due to chance occurrence of atypical combinations of stimulus configurations. In view of the number of subjects tested, however, this possibility is minimal.

In these analyses, the same subjects were included from previous analyses. New scores were developed for each subject based on the 60-trial test, and these scores were subjected to 4-way ANOVAs identical to those used for the analyses presented in Table 15. These analyses are summarized in Table 19, which compares the results with 60 trials to those with 80 trials.

It is obvious from this table that, while several of the variables continued to show essentially the same levels of significance as they did with 80 variables, there were some which were no longer significant. Most notably, variables from the matrix task showed no ability to differentiate among alcohol

**TABLE 19**

SIGNIFICANT ANOVA RESULTS — FAA-RTP RESCALED SCORES
BASED ON 60 TRIALS AS COMPARED TO 80 TRIALS

| FAA-RTP VARIABLE | ANOVA MAIN AND INTERACTION EFFECTS | | | |
|---|---|---|---|---|
| | SESSION | | SESSION X ALC/NON-ALC INTERACT. | |
| NO. OF TRIALS | 60 | 80 | 60 | 80 |
| T1RT | | | .0001 | .0001 |
| T1PC | | | | |
| T1TRRT | | | .0058 | .0001 |
| T1TRPC | | | | |
| T1TRTP | | | .0058 | .0001 |
| T1TP | | .0399 | .0001 | .0001 |
| T2RT | | .0001 | .0001 | |
| T2PC | | | | |
| T2TRRT | | | | |
| T2TRPC | | | | |
| T2TRTP | | .0305 | | .0001 |
| T2TP | | | | .0001 |
| ATTNRT | .0001 | .0001 | | .0326 |
| ATTNPC | | | | |
| ATTNTP | .0001 | .0001 | .0179 | .0145 |

TABLE 20

RELATIVE SENSITIVITY AND SPECIFICITY VALUES
FOR FAA-RTP USING BOTH
60 AND 80 STIMULUS PRESENTATIONS

A: CUT-SCORE: -2.0 STANDARD DEVIATIONS BELOW BASELINE
SCORING CRITERIA: FAIL ON ANY TWO FAA-RTP VARIABLES

| | NO. OF STIMULI | ALCOHOL LEVEL 0 | .02 | .04 | .06 | .08 |
|---|---|---|---|---|---|---|
| "PASS" | 80 | 58 | 9 | 10 | 5 | 1 |
| | 60 | 66 | 16 | 12 | 12 | 8 |
| "FAIL" | 80 | 25 | 32 | 31 | 35 | 38 |
| | 60 | 17 | 26 | 30 | 29 | 32 |
| SENSITIV | 80 | | 78% | 76% | 88% | 97% |
| | 60 | | 62% | 71% | 71% | 80% |
| SPECIF. | 80 | (Based on initial test on each day) = 70 | | | | |
| | 60 | (Based on initial test on each day) = 80 | | | | |

B: CUT-SCORE: -1.5 STANDARD DEVIATIONS BELOW BASELINE
SCORING CRITERIA: FAIL ON ANY TWO FAA-RTP VARIABLES
OR FAIL ON 6 SPECIFIC VARIABLES

| | NO. OF STIMULI | ALCOHOL LEVEL 0 | .02 | .04 | .06 | .08 |
|---|---|---|---|---|---|---|
| "PASS" | 80 | 46 | 3 | 0 | 4 | 0 |
| | 60 | | 9 | 7 | 7 | 3 |
| "FAIL" | 80 | 36 | 38 | 41 | 36 | 39 |
| | 60 | | 32 | 34 | 33 | 36 |
| % SENSITIV | 80 | | 92 | 100 | 90 | 100 |
| | 60 | 78 | 83 | 83 | 92 | |
| % SPECIF. | 80 | (Based on initial test on each day) = 56 | | | | |
| | | (Based on initial test on each day) = 61 | | | | |

conditions. The angles task, on the other hand, continued to show significant differences, and the vector task also demonstrated almost as much significance with 60 trials as it did with 80.

In general, the overall impression is that, while 60 trials did not produce as much discriminatory power as 80 trials, there was enough evidence of sensitivity to warrant further study. Therefore, the conservative and liberal "scoring algorithms" identified previously were applied to individual subjects' data to determine whether the use of 60 stimuli would reduce sensitivity and specificity. These data are presented in Table 20, which shows the relative values for both 80 and 60 stimuli.

Inspection of this table reveals a drop in the sensitivity of the test to the various levels of alcohol, ranging from 5 to 17%. In practical terms, this would mean that up to 7 more individual subjects from this sample could have "passed" FAA-RTP with some level of alcohol in their system if a 60-trial test were used. Some of these could have been at fairly high levels of alcohol. On the other hand, the level of false positives appears to go down with the smaller number of stimuli. This suggests that more stimuli cause failures in individuals due to non-alcohol-related causes (e.g., boredom) and, therefore, could actually be counter-productive in real implementations. This possibility must be considered and tested in subsequent studies.

For the moment, a viable hypothesis is that the FAA-RTP test could be reduced to a smaller number of stimuli, with a corresponding reduction in testing time. Since it is theoretically undesirable to reduce testing time beyond the point where a subject can "mobilize resources, even when degraded, there is obviously a limit to how short the test can be made. This limit, however, has not yet been fully defined.

# SECTION 4
# SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

The FAA is interested in determining whether Readiness-to-Perform (RTP) tests can significantly enhance the margin of safety in aviation. To this end, the FAA conducted a broad market survey of existing RTP tests, and eventually selected 1 particular test for further study. This was the NovaScan™ test procedure.

A plan was developed to produce an FAA-specific version of the NovaScan test (the "FAA-RTP test") and to subject it to a series of validation studies to determine its utility in the FAA environment. The present study constitutes the first of these efforts.

The basic design of the study called for a controlled amount of alcohol to be administered to subjects, and their performance assessed at various levels of blood alcohol. Seventy-seven male subjects were administered alcohol sufficient to raise their breath alcohol content (BrAC) by .02% per hour, up to a limit of .08% BrAC. They were tested at BrAC levels of 0, .02%, .04%, .06%, and .08%. On a separate day, "sham" alcohol drinks were administered to the subjects as a double-blind control.

Reaction time, percent correct, and measures of information processing were obtained from the FAA-RTP test. An average of 22.8 training trials on NovaScan were performed by subjects prior to entering into the experiment (range was from 12 to 44 trials for individuals). Subjects took between 16 and 17 training sessions before reaching a stable plateau or baseline on most variables. Test reliability was between .76 and .94 for most test measures.

Statistically significant effects of alcohol on the FAA-RTP test were found for all 3 individual test procedures. These were monotonically related to alcohol level of the ascending limb of the alcohol ingestion curve, but not on the descending limb. BrAC levels of .08% generally produced decrements in reaction time variables at or near 2 standard deviations from the subject's baseline performance. At .04% BrAC, decrements generally were in the range of 1.25 to 1.5 standard deviations from the subject's mean.

Individual analyses were made of each subject's performance related to alcohol level. A cut score of 2.0 standard deviations would have detected 97% of the subjects at .08% BrAC, 88% at .06% BrAC, and 76% at .04% BrAC. These detection rates were accompanied by a first-test false positive rate of 30%. A cut score of 1.5 standard deviations resulted in detection of 100% of subjects at .04% and .08% BrAC, and 90% at .06% BrAC, with a first-test false positive rate of 44%.

From the results of the present study, the following conclusions appear justified for this sample of subjects:

1. The NovaScan test developed for the FAA is sensitive to levels of alcohol in the range of .04% BrAC.
2. The median training time for subjects to reach plateau levels on the FAA NovaScan test is approximately 2.75 hours, with a minimum of 10 training sessions required. In this unselected sample of subjects, a training time of approximately 3.9 hours would have been required to assure that 90% of the subjects would have reached plateau.
3. The NovaScan test is not sensitive to the effects of 8 hours of work on a variety of visually and cognitively demanding tasks, with the levels of task workload used in this experiment.
4. Subject acceptance of NovaScan is high, as is subject belief in the efficiency of NovaScan in detecting performance decrement.

It is recommended that a criterion-based study be carried out, which would cross-validate the sensitivity of the FAA-RTP in detecting an alcohol stressor, and which also would establish the relationship between performance on this test and performance in a real-world environment.

## SECTION 5
## REFERENCES


Cahalan, D., Cisin, I. H., and Crossley, H. M. (1967). American drinking practices; A national study of drinking behavior and attitudes. New Brunswick, NJ: Ritgers Center of Alcohol Studies.

Computer Technology Associates, Inc. (1987). *FAA Air Traffic Control Operations Concepts Volume I: ATC Background and Analysis Methodology*. Englewood, CO.

Damos, D. L. (1991). Examining transfer of training using curve fitting: A second look. *International Journal of Aviation Psychology, 1(1)*, 73-85.

Gilliland, K., and Schlegel, R. E. (1993). Readiness to perform testing: A critical analysis of the concept and current practices. DOT/FAA/AM-93/13, FAA Office of Aviation Medicine, Washington, D.C. 20591.

Gregory, R. J. (1987). *Adult Intellectual Assessment*. Boston: Allyn and Bacon.

Jones, M. B., Kennedy, R. S., and Bittner, A. C. (1981). A video game for performance testing. *American Journal of Psychology, 94(1)*, 143-152.

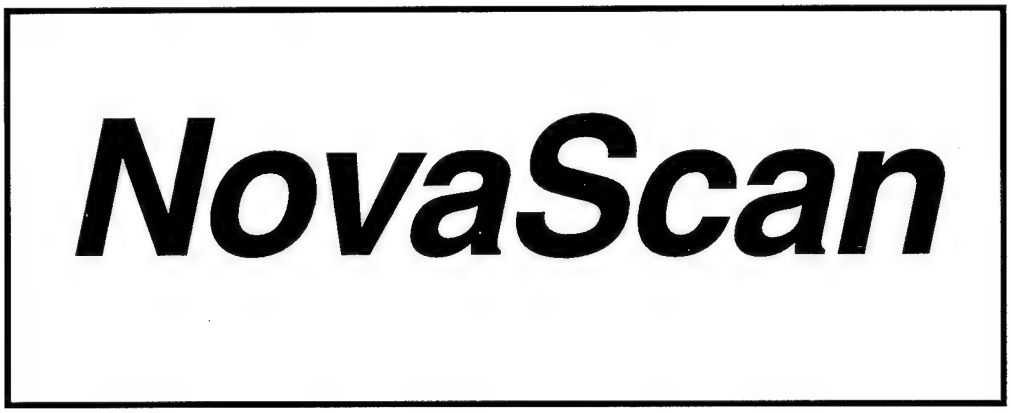
Lentz, S. K., and Rundell, O.H. (1976). Sustained control of blood alcohol levels. *Alcohol Technical Reports, 5(2)*, 33-36.

Lord, F. M., and Novick, M. (1968). *Statistical Theory of Mental Test Scores*. Wiley: New York.

Morrison, D.F. (1976). *Multivariate statistical methods*. New York: McGraw-Hill.

## APPENDIX A

## TUTORIAL ON THE FAA-RTP TEST



## NovaScan Instructions

### NovaScan Description

NovaScan consists of three tasks: (1) Continuous Spatial Memory, (2) Visual Search and Vector Projection, and (3) Attention Monitoring. The first two tests -- Continuous Spatial Memory and Visual Search and Vector Projection -- are presented, one at a time, in a randomly alternating fashion in the center of your screen. But the third task (Attention Monitoring) appears continuously on your screen in the form of a "frame" surrounding each of the first two tests. Combined, these three tasks measure your ability to find and remember items, form mental pictures of objects which change their positions, and detect rare events. Figure 1 shows examples of how your screen will appear.

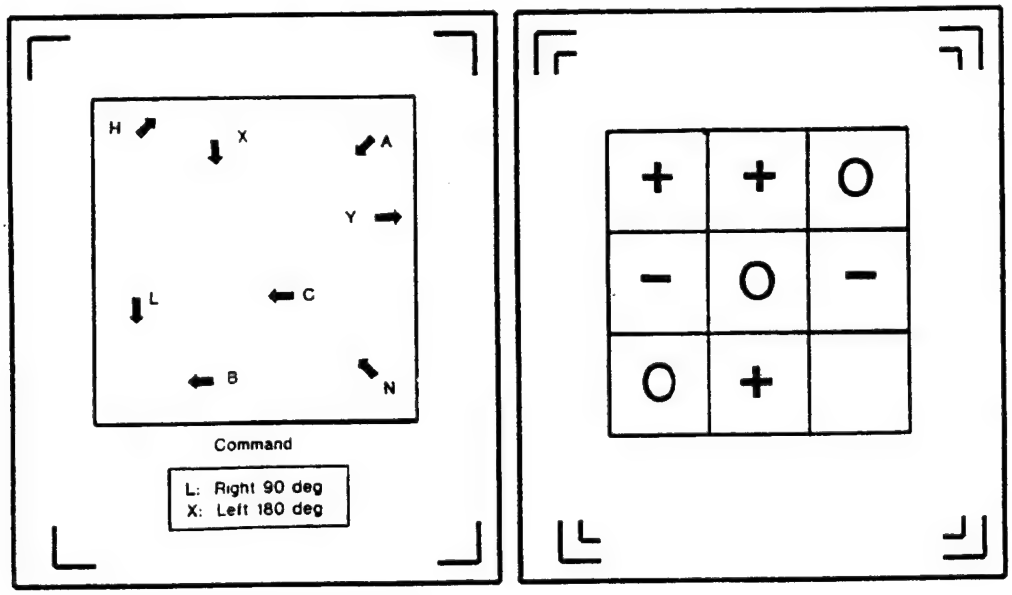



Figure 1. Sample NovaScan Test Screens

### Computer-Human Interface (CHI)

When you perform NovaScan, use the special keyboard provided, called the Computer/Human Interface, or "CHI", to respond to each of the tasks. The NovaScan CHI (see Figure 2) has a (1) keypad containing the numbers 0-9, plus "Enter" and "Backspace" keys; (2) two sets of red and green response keys; (3) two white thumb-keys; and (4) a joystick. [The

joystick, positioned in the middle of the CHI, will not be used for this version of NovaScan. Ignore the joystick throughout NovaScan testing.]

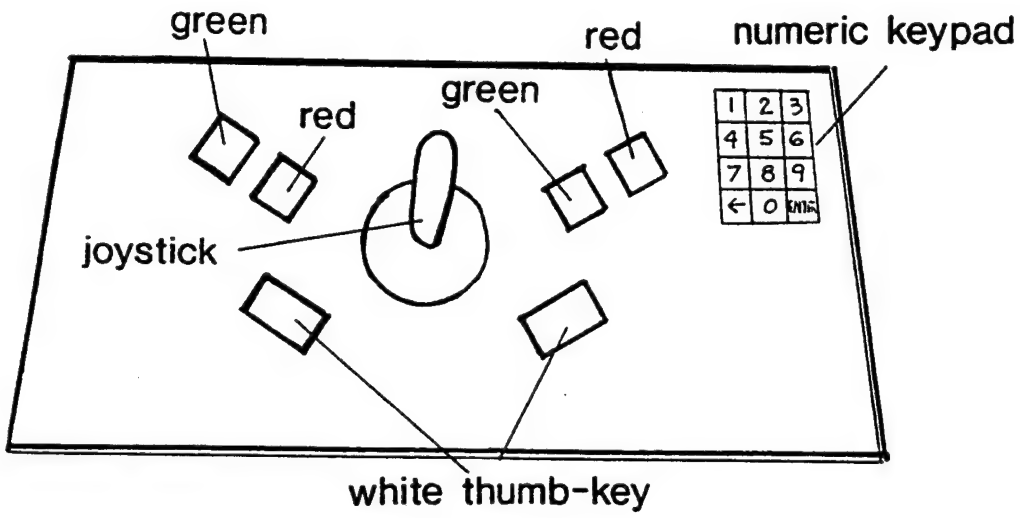



Figure 2. Computer/Human Interface ("CHI")

**Getting Started**

The following introductory screen (Figure 3) will appear at the beginning of each NovaScan test. To proceed, press any key on the CHI.

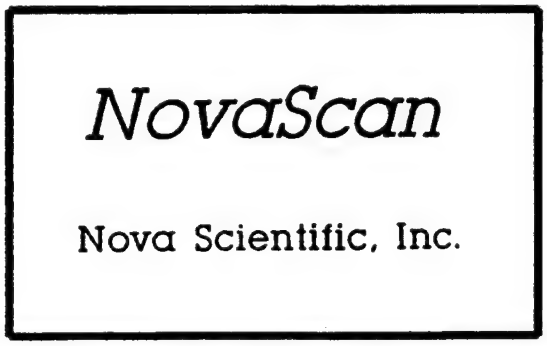



Figure 3. NovaScan Introductory Screen

Now, read the instructions on the Start-Up Screen (Figure 4). They tell you to enter your Social Security Number (SSN), using the keypad in the upper right corner of the CHI. If you make a mistake, use the "Backspace" key to back up and change the number. When you have correctly typed the number, press "Enter".

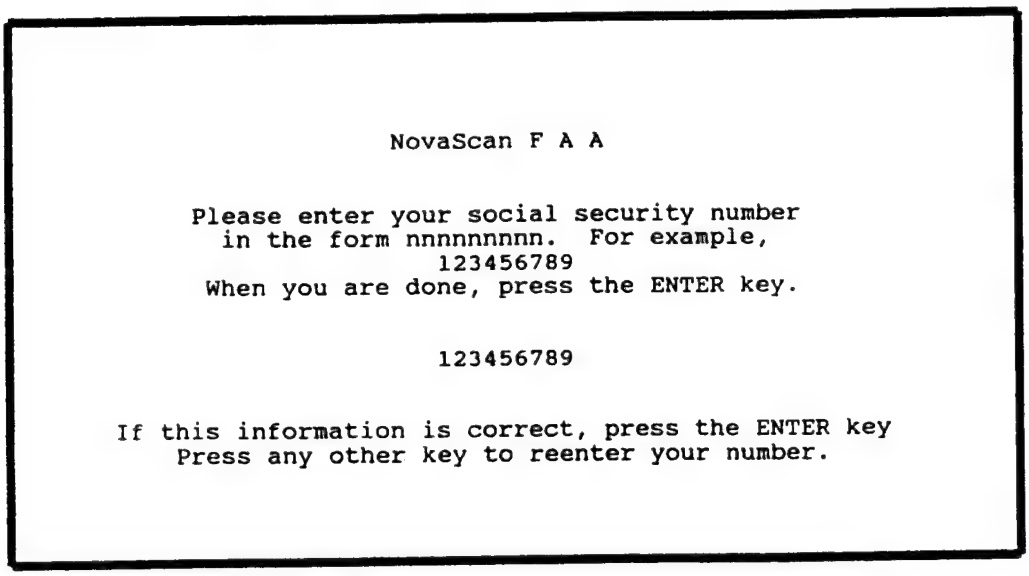



Figure 4. Start-Up Screen

Your SSN is your personal identification number for NovaScan, and assures that your NovaScan score is always recorded in your own directory. If the computer does not recognize your SSN, the following message will appear at the bottom of the screen: "This is not an active number, please try again". This message may mean that (1) you have mistakenly entered the wrong SSN, or that (2) you are seated at the wrong computer terminal. First, make sure that you are seated at the correct computer, then type your SSN again. If the computer still will not accept your data, please ask for help from one of the Study Administrators.

Next, place your fingers on the red, green, and white keys as follows (see Figure 5):

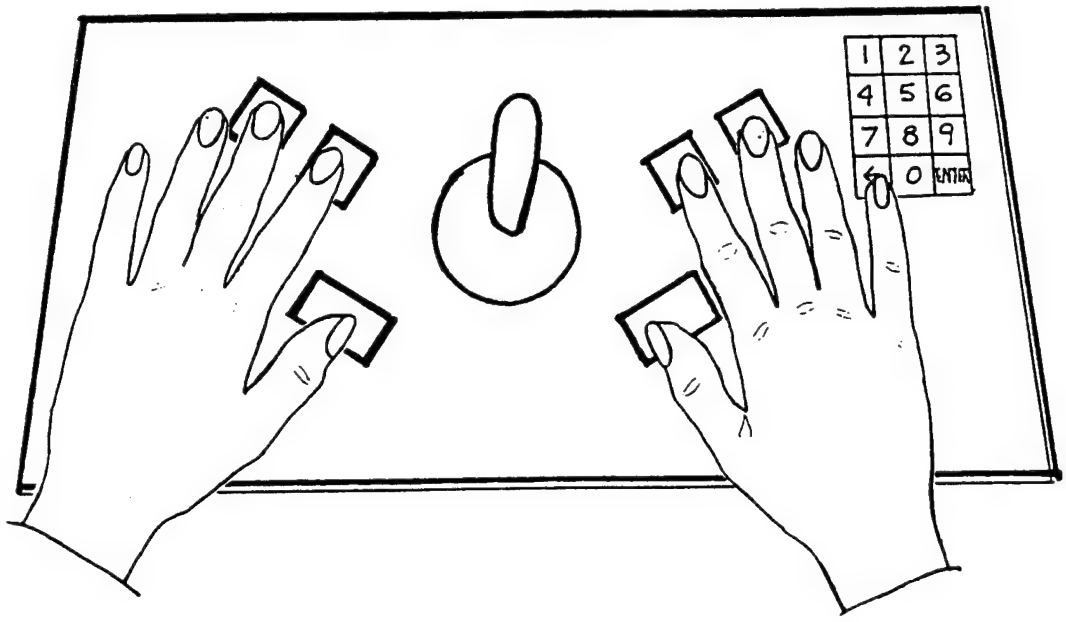



Figure 5. Appropriate Hand/Finger Placement on the Keys of the CHI

(1) Left hand: place your middle finger on the left (green) key, your index finger on the right (red) key, and your thumb on the white key.

(2) Right hand: place your index finger on the left (green) key, your middle finger on the right (red) key, and your thumb on the white key.

To assure your best scores, <u>always keep your hands and fingers in these positions during the entire NovaScan session every time you take NovaScan</u> -- poised and ready to respond rapidly.

**NovaScan Phases**

The NovaScan procedure has two phases: <u>Warm-up</u> and <u>Final</u>.

The <u>Warm-Up</u> phase is a shortened version of the Final phase. It lets you practice, or "warm up," before taking the final.

The Warm-Up phase calculates and displays your scores for your own information, but the computer does not record them.

The Final phase is a longer version of the Warm-Up. The final phase scores are the most important -- they are the ones that NovaScan uses to measure your performance for this experimental evaluation.

**Training**

NovaScan is a challenging test to learn. Remember that two of the tasks alternate on the screen randomly. Switching back and forth between tasks can make it very difficult to learn either one well. Therefore, to make it easier for you to learn NovaScan, you will train on each task individually first, to get started. Once you have learned each of Tasks 1 and 2, you may take the NovaScan test as it will be presented during the test phase of the study -- integrated into a randomly alternating pattern of test item presentation.

A special Training Program has been prepared for you. Your Trainer will help you select the appropriate steps to perform. This Training Program lets your Trainer select the task for you to practice (either Task 1, Task 2, or both) and also whether he or she wants you to receive feedback on each test item. In other words, the Training Program can be programmed to tell you if you responded incorrectly on any test item. This feedback will assist you in learning the correct way to respond to NovaScan. If your Trainer opts not to show you feedback, it is to give you an idea of how well you can perform NovaScan on your own.

When you begin training, a menu screen (Figure 6) will appear on your computer after you have entered your SSN. To begin, your Trainer will ask you to select Choice #1 (Visual Search and Vector Projection). Once you feel comfortable with this task, your Trainer will suggest that you return to the menu and select Choice #2 (Continuous Spatial Memory). When you have learned both tasks, your Trainer will ask you to select Choice #3 (Both Tasks). Choice #3 gives you an opportunity to practice NovaScan as it will appear during the actual experiment -- with both Tasks 1 and 2 alternating randomly on your screen.

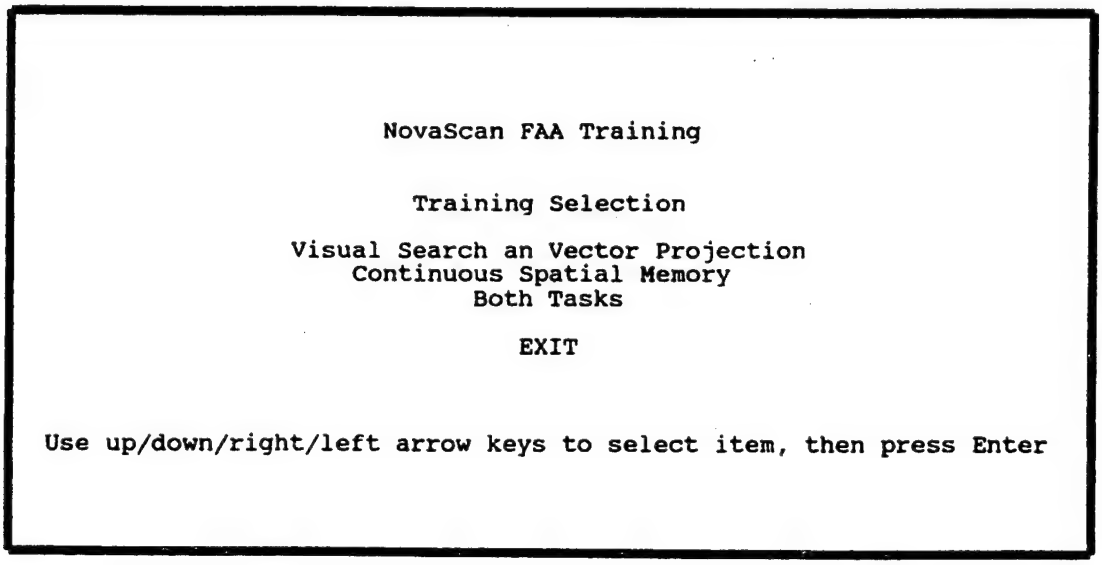



Figure 6. Menu Screen for Training

**Performing the Individual NovaScan Tasks**

Task 1: Continuous Spatial Memory

Task 1 measures your ability to detect, locate, and remember items. A 3 X 3 matrix will appear on your screen which contains three types of symbols: "+", "-", and "o". There will always be three each of two of these symbols, but only two of the remaining symbol. These eight symbols will be positioned randomly within the 3 X 3 matrix. One square will remain empty. See Figure 7, below.

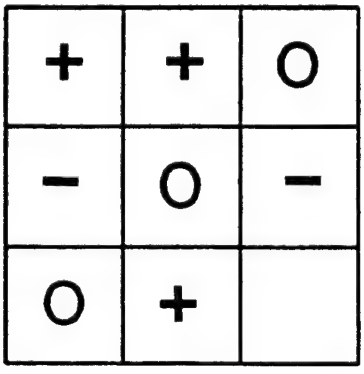



Figure 7. Example of Task 1

To perform this task, first scan the matrix. Determine the type of symbol which is missing and its location within the matrix, and remember them. Next, determine whether the missing symbol is both the same type of symbol missing from the same cell of the matrix as the previous time Task 1 was presented. If so, press the green key with the index finger of your right hand to respond "same". If not, press the red key with the middle finger of your right hand, for "different" (Figure 8).

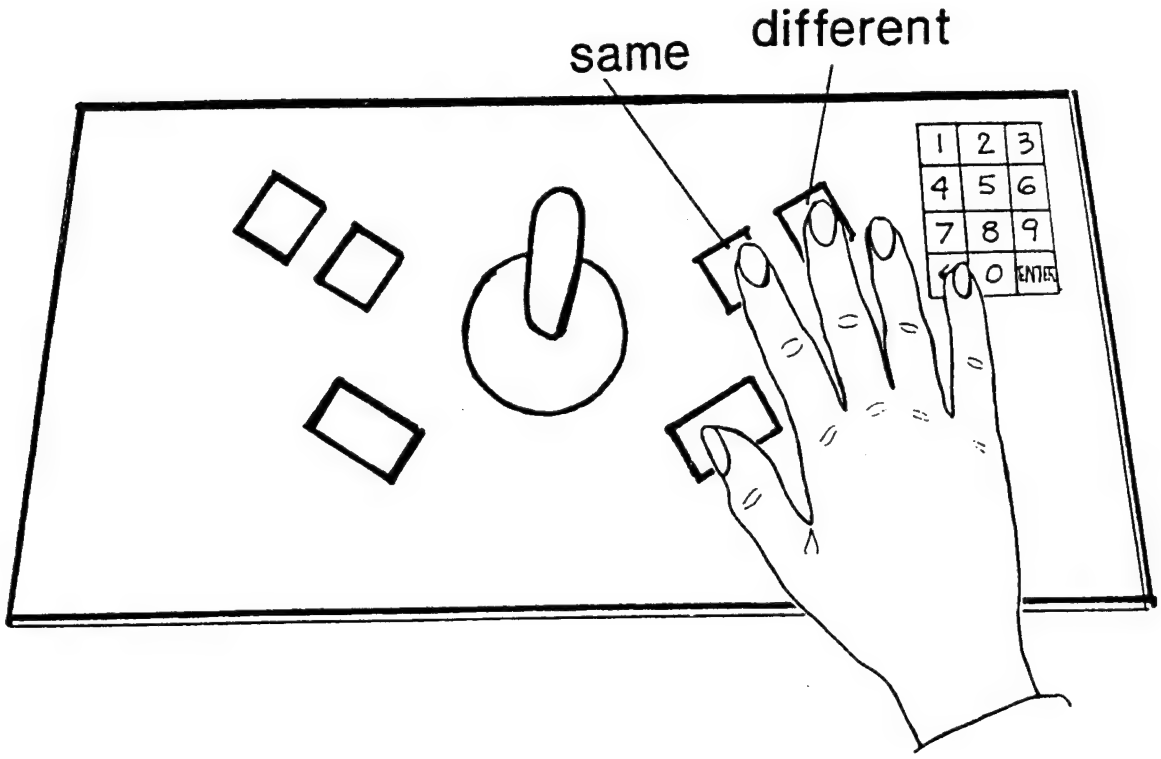
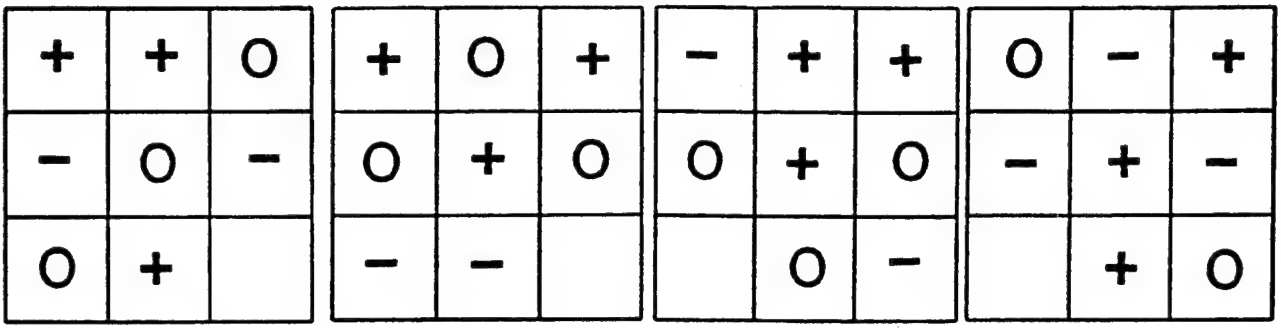



Figure 8. "Same" and "Different" Keys for Task 1

Note: The very first time Task 1 is presented on each NovaScan test, simply note and remember the missing symbol and its location, but respond with either "same" or "different", since there is no previous test item with which to compare it.

Correct responses in this sequence of sample tasks are (Figure 9):




Same/Different 1. Same 2. Different 3. Different 4.

Figure 9: Samples of Task 1 in Order of Their Appearance During NovaScan and Their Correct Responses

Notice that your response will be "different" unless both the missing symbol and location are identical to the previous test item for Task 1.

## Task 2: Visual Search and Vector Projection

Task 2 measures your ability to form mental pictures of objects -- in this case, arrows -- which are "instructed" by turn commands to change their position. Eight arrows are randomly positioned on your computer screen. Each arrow is pointed in any of eight directions: 0, 45, 90, 135, 180, 225, 270 or 315 compass degrees (Figure 10).

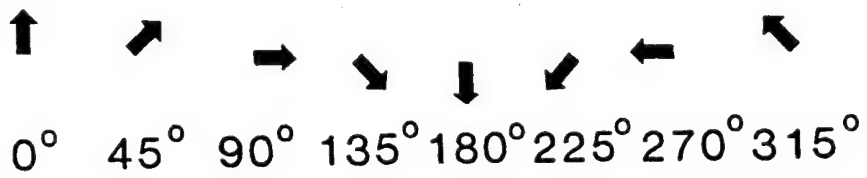



Figure 10. Eight Possible Positions for Task 2 Arrows

Each of the eight arrows is designated by a different letter of the alphabet. So, your computer screen may look something like this (Figure 11):

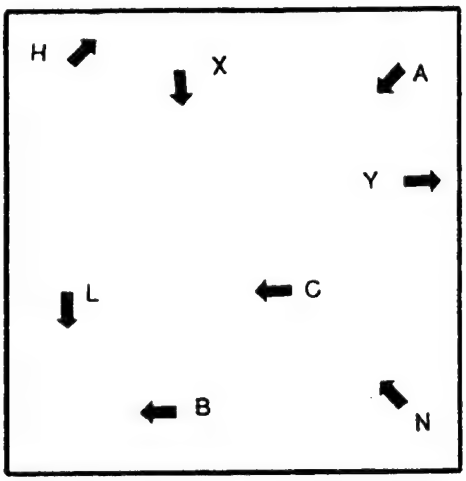



Figure 11. Sample Task 2

Below this array of arrows are two command lines. They describe which way <u>two</u> of the eight arrows have been "instructed" to <u>turn</u>. For example, the turn command lines might read like this:

A: Right 90 degrees
X: Left 180 degrees

These turn commands mean that the arrow labeled "A" would turn clockwise 90 degrees from its originating position and that the arrow labeled "X" would turn counter-clockwise 180 degrees from its originating position. It is your job to mentally turn these arrows the direction and amount stated in the turn command, and to mentally project the arrows into space across the display on your screen, in order to determine whether they would ever cross, or "conflict".

In the above example, the resulting positions of these two arrows would be as follows (Figure 12). Given their point of origin, the direction in which they are presented, and their turn commands, these arrows would conflict.

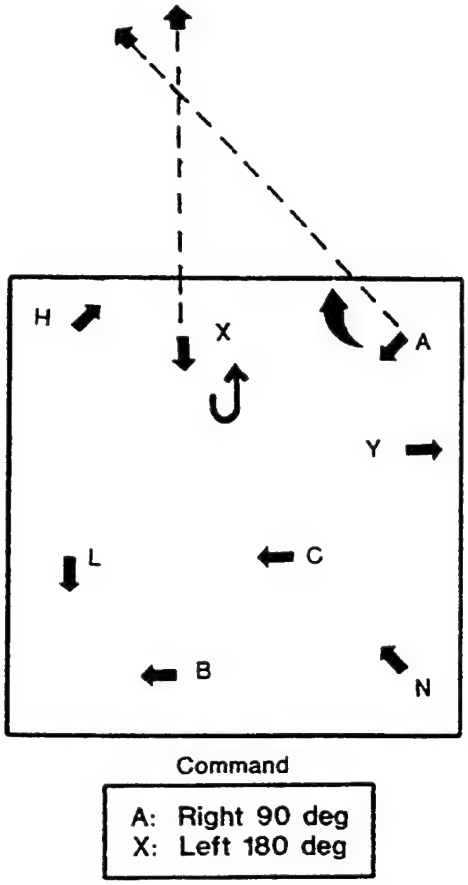



Figure 12. Resulting Positions for Figure 11 Arrows

Remember: A command to turn the arrow RIGHT means to turn it clockwise (Figure 13a). A command to turn the arrow LEFT means to turn it counter-clockwise (Figure 13b).

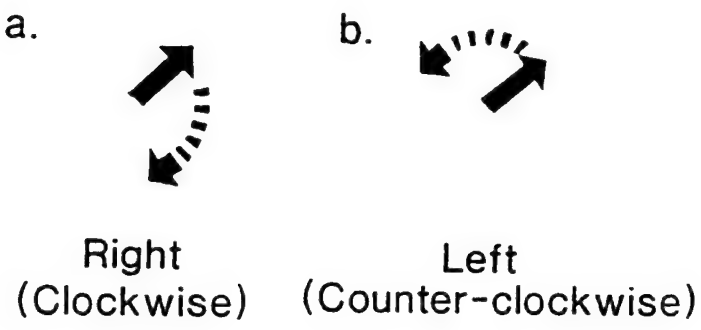



Figure 13. Direction of Arrow Movement for RIGHT and LEFT Turn Commands

If you determine that there would be a conflict between two arrows, respond with your left index finger by pressing the right (red) key. If the arrows would never cross, respond by pressing the left (green) key with your left middle finger (see Figure 14).

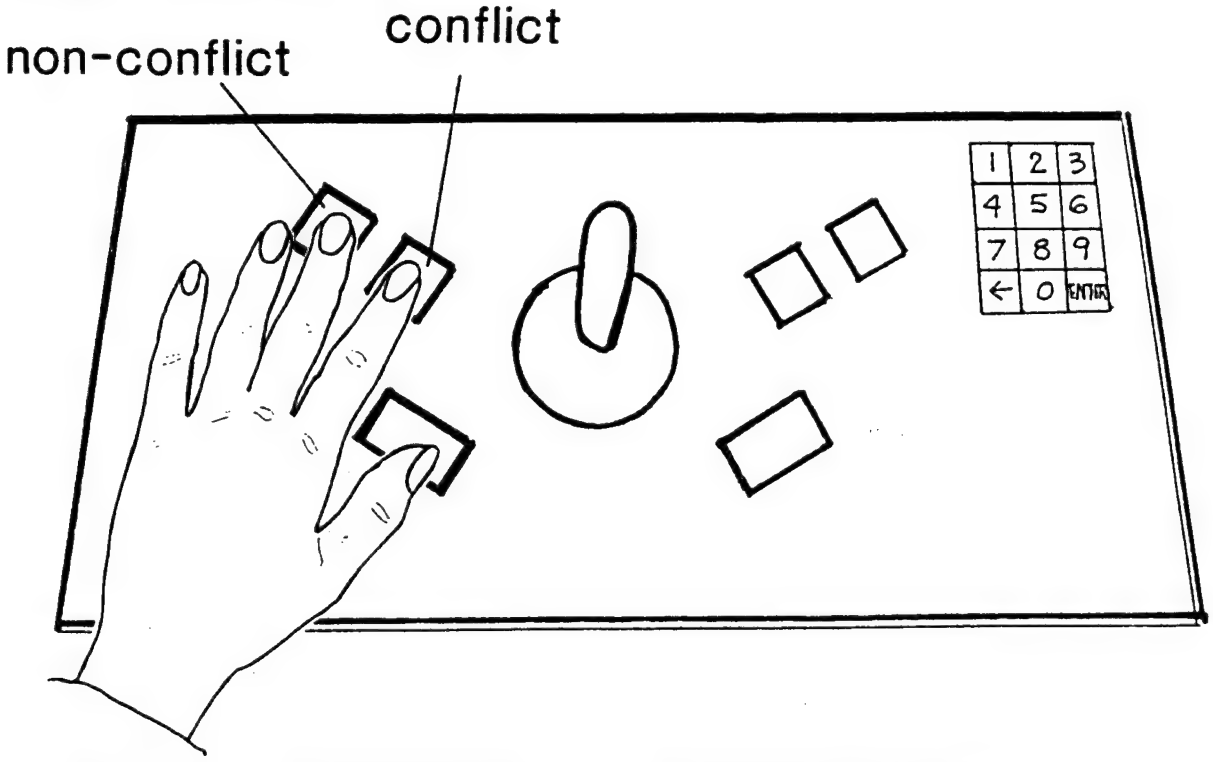



Figure 14. "Conflict" and "Non-conflict" Key Responses for Task 2

Task 3: Embedded Attention Monitoring

Task 3 monitors your ability to detect rare events. It is embedded in each of Tasks 1 and 2 in the frame surrounding each task. The frame looks like this (Figure 15):

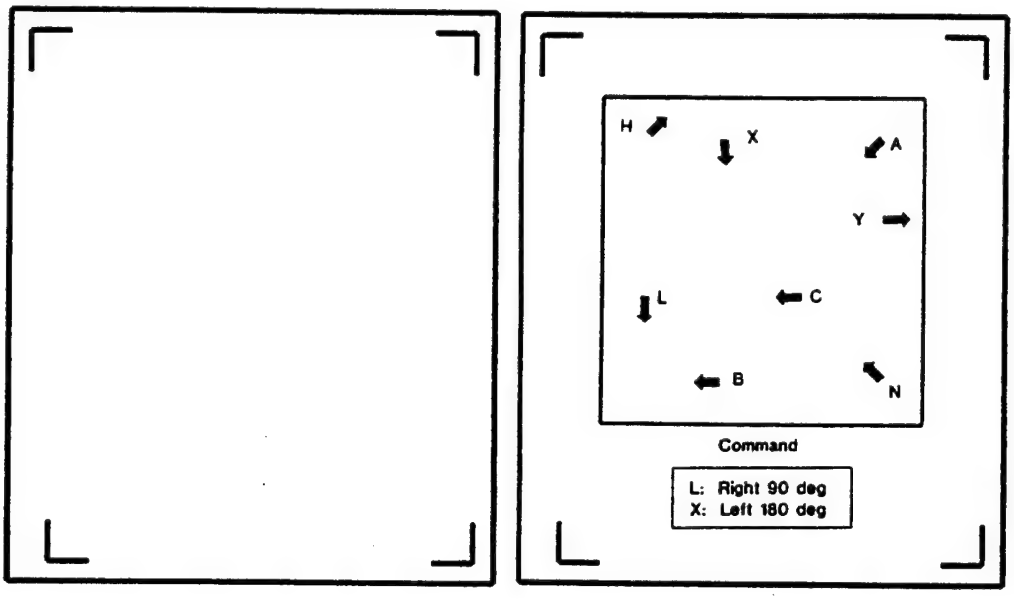



Figure 15. Embedded Task 3: A Frame Surrounds Each Task and Contains Angles in Each Corner

Notice the angles, or carrots, located inside each corner of the frame. Typically, there will be only single carrots in each corner. However, as each new task is presented, these angles may change to double carrots, as shown in Figure 16.

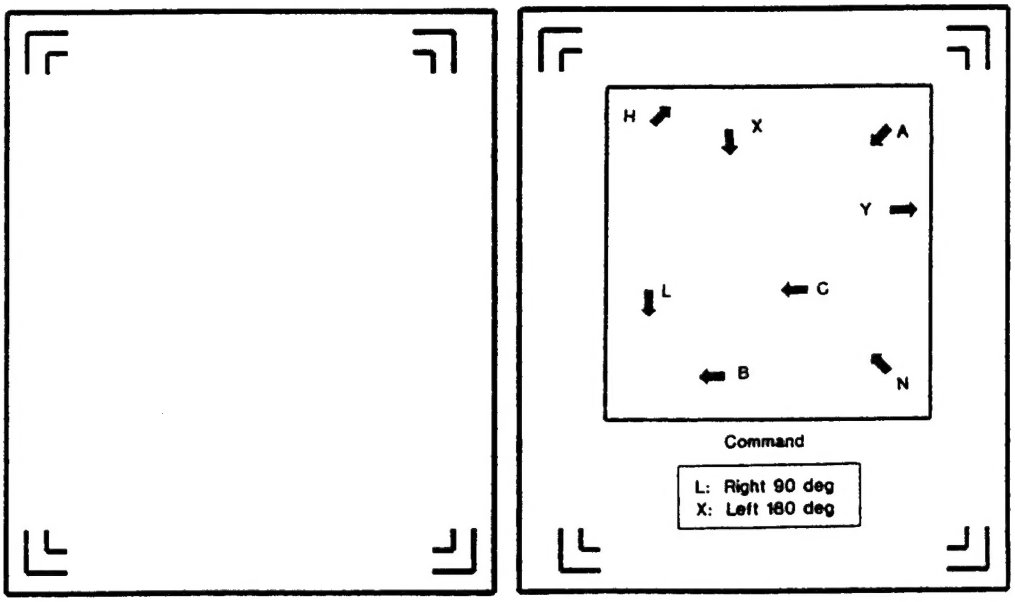



Figure 16. Embedded Task 3 with Double Carrots/Angles

When the carrots change from single to double ones, immediately press one of the two white thumb keys (Figure 17). (It does not matter whether you press the right or left white thumb-key.)

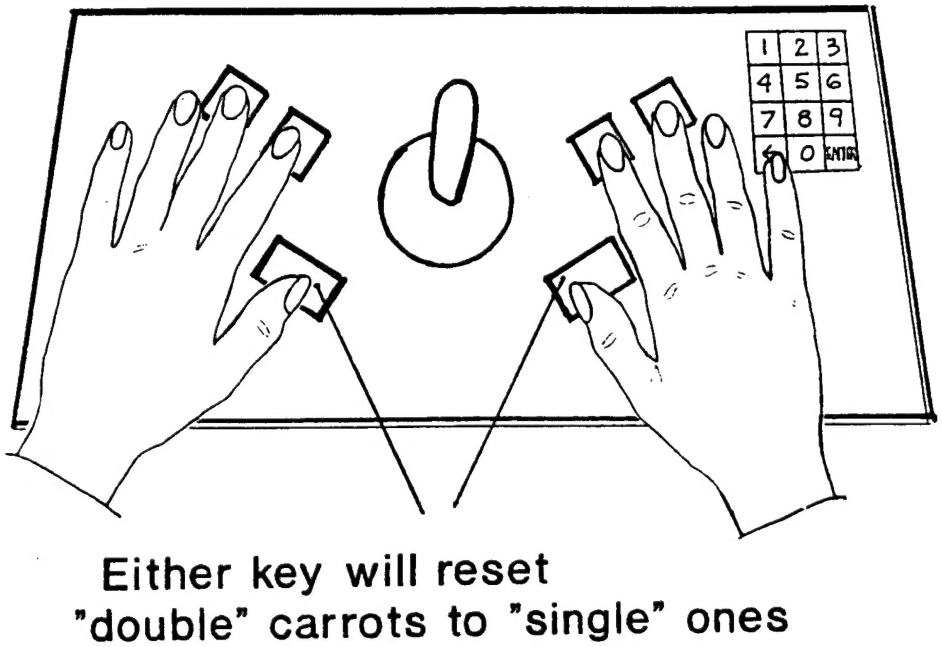



Figure 17. Use Either White Thumb-Key to Reset the Double Carrots to Single Carrots

**Remember:** Always respond <u>first</u> to the embedded task, <u>then</u> to Task 1 or 2.

<u>Warm-Up Phase</u>

Recall that the Warm-Up phase is a shortened, practice, version of the Final phase. At the end of each Warm-Up, the following results will appear for you to review (Figure 18).

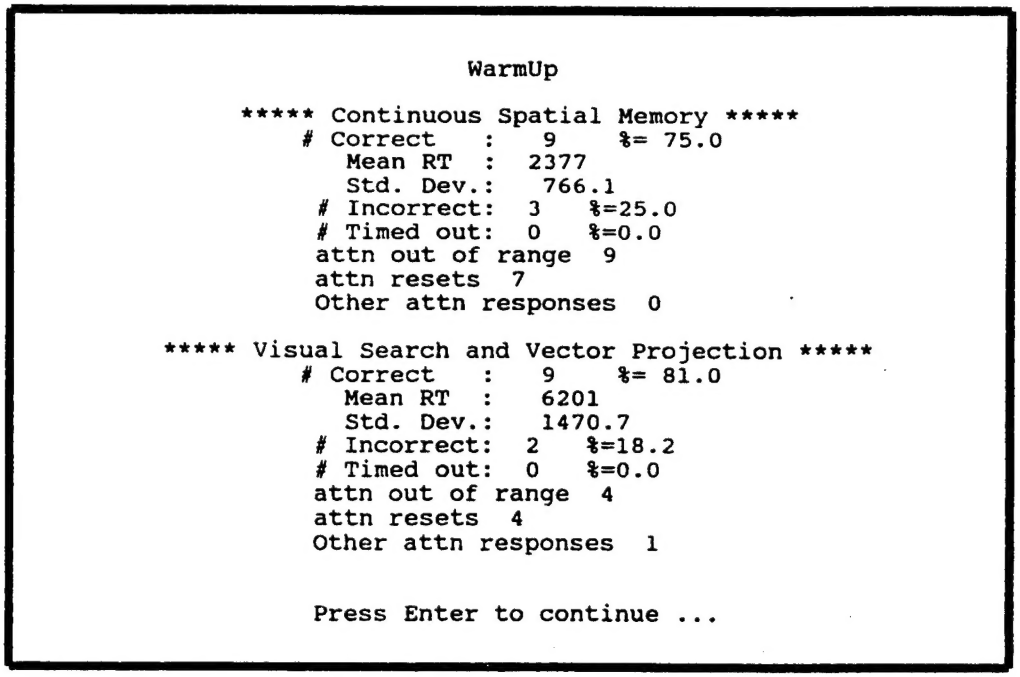



Figure 18. Warm-Up Results

For each of Tasks 1 and 2, "Continuous Spatial Memory" and "Visual Search and Vector Projection", you are provided with

the following scores: (1) number and percent correct, (2) mean reaction time (RT), (3) standard deviation (Std. Dev.) of your reaction time, (4) number and percent incorrect, and (5) number of stimuli which timed out. Your reaction times and standard deviations are expressed in milliseconds (e.g., 2129 is equivalent to 2.129 seconds).

"Standard deviation" describes how much your reaction time scores bounce around or vary from one test item to the next; the lower the standard deviation becomes, the more consistent or alike your response times have become. "Time-outs" are those test items which you did not respond to within the allotted time; so, the computer timed-out and proceeded to the next test item. As you improve, percent correct should increase, and mean reaction time, standard deviation, and number incorrect and timed out should decrease.

Additionally, the Warm-Up results display information as to how you performed on the embedded task (Task 3). The last three items in the Results section for each task -- "Attn out of range", "attn resets", and "Other attn responses" -- pertain to the embedded task.

"Attn out of range" indicates the number of times the carrots changed from single to double ones, and "Attn resets" indicates the number of times you correctly pressed the white thumb-key in response. If these two numbers are identical in the Results section, you responded perfectly on the embedded task throughout the NovaScan Warm-Up each time the carrots changed from single to double.

The last item, "Other attn responses", lists the number of times you inappropriately pressed the white thumb-key -- such as when the carrots were only single (rather than double).

These Warm-Up results are not saved; they merely give you an indication of how you have performed on the Warm-Up.

## Final Phase

Following each Warm-Up is the actual test phase of NovaScan. All data from this phase is automatically saved in your individual directory. For experimental purposes, however, all NovaScan Final phase data will not be shown to you, but will be saved in the computer for analysis following the study.

**Remember:** Perform as accurately as you can, and while doing so, move as quickly through the test items as you are able. Do not be discouraged if NovaScan is difficult to perform at first. It is a test of high intellectual

abilities, and everyone who takes it is challenged, particularly in the beginning.

**If You Have Questions**

If you have any questions, please ask your Test Administrator to help you. He or she will be happy to assist.



☆U.S. GOVERNMENT PRINTING OFFICE: 1995 - 660-067/20021